University of Alberta

EVALUATION OF A THREE-LEVEL FERAM

by

Kamlesh Ram Raiter

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science**

Department of Electrical and Computer Engineering

Edmonton, Alberta
Spring 2005

0-494-08160-0

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file *Votre référence*
ISBN:
Our file *Notre reference*
ISBN:

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

...Still round the corner there may wait
A new road or a secret gate;
And though I oft have passed them by,
A day will come at last when I
Shall take the hidden paths that run
West of the Moon, East of the Sun.

Frodo's Song
"The Grey Havens"
Return of the King
- by J.R.R. Tolkien

To my parents: Meera and Ram Raiter

Abstract

Ferroelectric random-access memories (FeRAMs) face a big challenge in terms of larger cell area and larger chip area because of the drivers and wiring required for the platelines. This research proposes an idea of multilevel storage (3 levels per cell) in FeRAM, similar to multilevel storage in flash memories. This work addresses some of the major challenges in implementing such multilevel storage. In particular read and write mechanisms for such a multi-level cell are investigated. A parameter sensitivity analysis is done to demonstrate the feasibility of three-level FeRAM operation in the presence of inevitable parameter variations in the cell array. Probability distribution function plots are presented for studying the voltage distribution of the various signal levels. Noise margins of ~100mV are achieved between the signal levels with readout voltage variations of around 20-30%. Three-level data signaling increases the storage density for 1T1C cells by up to 50%. Issues relating to stable cell operation, along with technological and design constraints are discussed. Finally, some techniques to overcome these issues are proposed like auto-calibrated reference voltage generation scheme and damped oscillating wave bitline voltage driver.

Acknowledgements

I would like to show my appreciation towards my supervisor Dr. Bruce F. Cockburn for his valuable guidance and support throughout the course of this research. I admire him most for his enthusiasm and tireless efforts as a mentor. I would like to thank Dr. Igor Filanovosky, Kris Breen, Christian Giassion, Tyler Brandon, John Koob, Jesus Tapia and all my other fellow graduate students in the VLSI lab for their patience in understanding my problems and for all the invaluable technical support that they extended to me over the duration of my project. I would like to thank Paul Greidanus and Jacob Bresciani for providing me help with the hardware and software in the VLSI lab. I would also like to thank Sanjeeva Srivastava and Vijaya Raghavan for their inspiration and motivation during my masters at the University of Alberta. I would like to thank Dr. Vincent Gaudet and Dr. Mike MacGregor for being in my examining committee and providing valuable feedback. Also, I would like to thank NSERC for financial support and the Canadian Microelectronics Corporation for providing hardware and software tools on which this project was mostly carried out.

Table of Contents

Chapter 1 Introduction 1

1.1 Motivation 1

1.2 Objectives 5

1.3 Thesis Outline 6

Chapter 2 Background 9

2.1 Ferroelectricity 9

2.2 The Ferroelectric Capacitor (FeCap) 12

2.3 Basic Memory Cell Structure 13

2.3.1 DRAM-like FeRAM cells 13

2.3.2 MFSFET FeRAM Cell 14

2.4 Two-level Operation 15

2.4.1 Read and Write Sequences 15

2.4.2 Sensing and Amplification 18

2.5 FeRAM Design Challenges 20

2.5.1 Reference Voltage Generation 20

2.5.2 Access Time Delay 21

2.5.3 Reliability Issues 24

2.5.4 Memory Densities 26

2.6 FeRAM Architectures 27

Chapter 3 Review of Multilevel Memory Technologies 33

3.1 Overview 33

3.2 Multilevel Flash Memory 34

3.2.1 Programming Techniques in ML Flash 35

3.2.2 ML Sensing Schemes 37

3.2.3 Reliability Issues 38

3.3 Multilevel DRAM 39

3.3.1 Writing Schemes 40

3.3.2 Sensing and Reference Generation Schemes 40

3.4 MLDRAM Challenges 42
3.5 Summary 43
Chapter 4 Multilevel Ferroelectric Memory Design 45
4.1 Multilevel Cell Design 46
4.2 Read and Write Sequences 50
4.3 Design of the Middle Level Voltage Driver 53
4.3.1 Method I 54
4.3.2 Method II 63
4.4 Reference Voltage Generation 69
4.4.1 Memory Core Floorplan 69
4.4.2 Array Architecture 70
4.4.3 Error Checking/Auto Calibration Mode 74
Chapter 5 MLFeRAM Operation in the Presence of Parameter Variations 77
5.1 Schematic Simulation Setup 79
5.2 Review of the Normal Distribution 80
5.3 Parameter Variations 81
5.3.1 Process Variations 81
5.3.2 Simulation Technique for Noise Margin Analysis 84
5.3.3 Variations in V_M 94
5.3.4 Implementation Alternatives and Cell Signal Distributions 97
5.3.5 Write Sequence Variations 102
Chapter 6 Conclusions 105
6.1 Summary 105
6.2 Challenges 110
6.3 Suggestions and Future Work 112
Bibliography 115
Appendix A 121
A.1 121
A.2 123
A.3 127
Appendix B 129

Appendix C 136
Area Estimation 136

List of Tables

Table I Characteristics of Emerging Non-volatile Memories [20] [53] [54] 4

Table II SPICE BSIMV3 Model Parameters for the CMOS 0.35-μm TSMC Process 61

Table III Operational Amplifier Characteristics 67

Table IV Spectrum of the Values of V_{REF1} and V_{REF2} 73

Table V Noise Margins for Variations in V_M 94

Table VI Readout Voltages (V_R) for Type I and Type II Pulsing Sequence 103

Table VII Readout Voltages after Writing a Sequence of Three Voltages 103

Table VIII FeRAM and Chain FeRAM Chip Size Comparisons 136

List of Figures

Figure 1-1 Forecast for Memory Market Demand in Billions of Dollars [Adapted from Web-Feet Research, Inc.] 2
Figure 2-1 Dual-state Ferroelectric Lattice Under the Influence of Positively and Negatively Oriented Electric Fields [48] 10
Figure 2-2 (a) Hysteresis Loop (Plot of Polarization vs. Electric Field) (b) Domain Changes in Response to Applied Electric Fields. 11
Figure 2-3 (a) 1T-1C cell (b) 2T-2C cell 13
Figure 2-4 Ferroelectric Memory Cell Structure [25] 14
Figure 2-5 (a) Write Waveforms (b) Polarization States During FeRAM Write Operations 16
Figure 2-6 (a) Read Waveforms (b) Polarization States During FeRAM Read Operations 18
Figure 2-7 Non-driven PL Scheme 22
Figure 2-8 Read Operation in the Differential Capacitance Read Scheme [14] 23
Figure 2-9 Comparison of Read Access Times [14] 24
Figure 2-10 Effect of (a) Fatigue, (b) Imprint and (c) Relaxation 26
Figure 2-11 The (a) WL||PL and (b) BL||PL architectures [1] 28
Figure 2-12 Merged Plateline Architecture [1] 28
Figure 2-13 1T-2C Cell Structure 29
Figure 2-14 Chain FeRAM 30
Figure 3-1 Floating Gate Memory (Flash) 34
Figure 3-2 Four-level Signaling (2 bits per cell) 35
Figure 4-1 Introducing a Third State 46
Figure 4-2 Charge Transferred to BL due to Switching and Non-switching Charge 47
Figure 4-3 Thevenin Equivalent Circuit for the Switched Charge 47
Figure 4-4 Ferroelectric Capacitor Characteristics [12] 51
Figure 4-5 WL, BL and PL Waveforms for Write Operations. 52

Figure 4-6 Hysteresis Trajectories During ML Write Operations 52
Figure 4-7 Hysteresis Trajectories During ML Read Operations 53
Figure 4-8 Two-stage Operational Amplifier 55
Figure 4-9 Block Diagram of OPAMP as Voltage Follower 57
Figure 4-10 Depolarizing the Material to the Zero Polarization State 59
Figure 4-11 Reaching the Zero Polarization State Through Spiral Inner Subloops 59
Figure 4-12 Write Control Waveforms 60
Figure 4-13 Transient Response for a Step Input 62
Figure 4-14 Writing V_M with (a) Differentially Pulsing BL and PL (b) Constant PL Voltage 66
Figure 4-15 (a) Current Steering DAC (b) Single Transistor Current Source 67
Figure 4-16 Input-output Characteristics of the 5-bit Current Steering DAC 68
Figure 4-17 Memory Core Floorplan 70
Figure 4-18 Architecture of an Array Section 71
Figure 4-19 Internal Structure of the Precharge Block 71
Figure 4-20 Bath-tub Curve Obtained by Sweeping the Reference Voltages 75
Figure 5-1 Cell Charge Distributions for Data 0 and Data 1 [40] 79
Figure 5-2 (a) Probability Distribution Function; (b) Cumulative Distribution Function 81
Figure 5-3 Readout Voltages with Variations in (a) Area and (b) C_{BL} 82
Figure 5-4 Readout Voltages with Variations in (a)ε_r and (b) Ps 83
Figure 5-5 Normal Distributions for the Readout Voltages 86
Figure 5-6 Variations in V_H, V_M and V_L with Variations in the FeCap area 88
Figure 5-7 Variations in V_H, V_M and V_L with Variations in the Bitline Capacitance C_{BL} 89
Figure 5-8 Variations in V_H, V_M and V_L with Variations in the C_{BL}/C_{FE} ratio 90
Figure 5-9 Readout Voltage vs. Cell Area and BL Capacitance 91
Figure 5-10 Effects of 10% (left) and 20% (right) Standard Deviations in Cell Areas of (Top) 0.25μm^2 cell (Middle) 0.65μm^2 cell (Bottom) 1μm^2 cell [X-Axis = Readout Voltage in mV, Y-Axis PDF] 92

Figure 5-11 Cell Signal Distributions for Cell Sizes (a) $2\mu m^2$ (b) $3\mu m^2$ 93
Figure 5-12 Distributions with Variation in V_M (a) Minus 5% (b) Plus 5% (c) Minus 10% (d) Plus 10% (e) Minus 15%(f) Plus 15% [X-Axis = Readout Voltage in mV, Y-Axis PDF] 97
Figure 5-13 Cell Signal Distributions with Implementation Alternatives: (a) Original (b) Damped Transient from the OPAMP (c) Damping Wave Generated from DAC 98
Figure 5-14 Reduction in the Spread of the Distribution of V_M with Various Techniques (DAC, OPAMP, VPWL) 99
Figure 5-15 Noise Margin Comparisons for Alternative Sensing Techniques ... 100
Figure 5-16 Cell Signal Distributions With the After-pulse Sensing Technique: (a) 20% and (b) 40% Variation 101
Figure 5-17 Pulsing Techniques: (a) Type I, (b) Type II 102
Figure 6-1 Voltage Distribution due to Overall Parametric Variations [22] Log-scale 107
Figure 6-2 Voltage Distribution due to Overall Parametric Variations [22] Linear-scale 107
Figure 6-3 Voltage Distribution for a Three-level FeRAM with Variation in Cell Area 108

List of Abbreviations

1T-1C	One Transistor One Capacitor
ADC	Analog-to-Digital Converter
BL/BLN	True Bitline/Complementary Bitline
C_{BL}	Capacitance of the Bitline
C_{FE}	Capacitance of the Ferroelectric Capacitor (non-switching)
CMC	Canadian Microelectronic Corporation
CMRR	Common Mode Rejection Ratio
C_{OX}	Oxide Capacitance
C_{PL}	Capacitance of the Plateline
DAC	Digital-to-Analog Converter
DRAM	Dynamic Random-Access Memory
F	Minimum feature size for a given technology
FeCap	Ferroelectric Capacitor
FeRAM[1] FRAM™	Ferroelectric Random-Access Memory
ML	Multi-level
MLDRAM	Multilevel Dynamic Random-Access Memory
ML Flash	Multilevel Flash
μn	Mobility of the negative charge carriers (electrons)
NDRO	Non-destructive readout
NVM	Nonvolatile Memories
OPAMP	Operational Amplifier
PL	Plateline
Pr	Remnant Polarization

[1] FRAM is a trademark of Ramtron International Corporation (Colorado Springs, CO, USA).

P_S	Saturation Polarization
PSRR	Power Supply Rejection Ratio
Q_r	Charge Corresponding to the Remnant Polarization
SA	Sense Amplifier
V_c	Coercive Voltage
V_{DD}	Power Supply Voltage and Source of Logic High
V_{ref}	Reference Voltage
V_{SS}	Power Supply Voltage and Source of Logic Low
WL	Wordline

Chapter 1

Introduction

1.1 Motivation

Nonvolatile memories (NVM) play a significant role in many digital systems. NVMs are widely used in embedded controllers and consumer products, such as cellular telephones, for storing program instructions and other fixed data. They are especially important for certain applications, such as medical, military and space systems, where vital information has to be stored in a memory that does not lose the stored information when the power is lost. Contactless smart cards require nonvolatile memories with low power consumption as the cards use only wireless electromagnetic coupling to power up the electronic chips on the card. This application, unlike many other NVM applications, cannot use battery-backed up volatile static random-access memory (SRAM). Digital cameras require both low power consumption and frequent fast writes in order to read and/or write an entire image into the NVM in less than one tenth of a second [1]. Among the available nonvolatile memory technologies, disk memories offer largest storage capacities at the cheapest price per bit, and are widely used in such items as file servers and personal computers. However, they are slow, bulky, and susceptible to breakdown because of their mechanical nature. Magnetic core memory and

magnetoinductive plated wire memories are limited in capacity, are bulky, have large power requirements and are expensive (per bit). Core memory has therefore disappeared from modern computer systems.

NVM envisaged a phenomenal growth since the introduction of *floating gate memories*[2] in 1970s. Also called erasable programmable read-only memory (EPROM), hot electrons are injected into the floating gate by applying a lateral electric field between the source and drain and they (electrons) are removed by exposing the memory to ultraviolet rays or by means of Fowler-Nordheim tunneling [25]. An amenable alternative to EPROM has been electrically erasable PROM (EEPROM), which offers in-circuit programmability and erasability. Flash memories are comparatively new and are based on either EPROM or EEPROM technologies, in which the contents of memory cells across the array are erased simultaneously. Since Flash memory cells do not need individual erasure of memory cells they are two to three times smaller than EPROM cells.

During the last decade the use of NVM has increased dramatically because of their pivotal role in upcoming new applications like mobile and handheld devices. Figure 1-1 shows the projected memory market demand for the next few

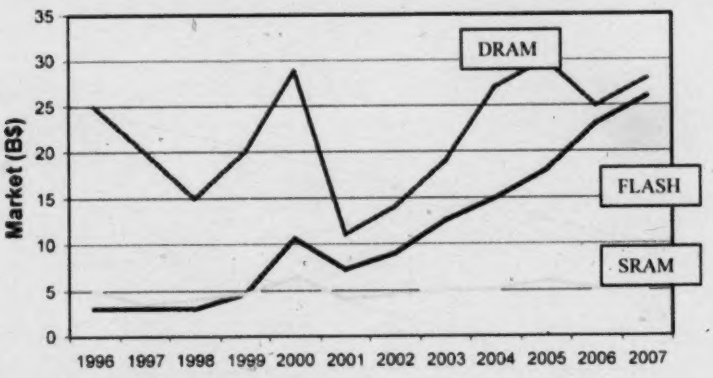


Figure 1-1 Forecast for Memory Market Demand in Billions of Dollars [Adapted from Web-Feet Research, Inc.]

[2] Charge (electrons) is stored on the floating gate leading to logic '0' for presence of electrons and logic '1' otherwise.

years. This increase in their applicability for wide range of applications imposes a further demand on their performance, density and reliability characteristics. In the last few decades Flash memory became such a dominant NVM type that the term "flash" has almost become synonymous with the term "non-volatile memory". This was due to its scalability and the high achievable storage capacities (now exceeding 1Gb). However, technical difficulties are expected to make further scaling of Flash memory increasingly difficult. Also, Flash memory has slow write times, requires relatively high operating voltages, is susceptible to radiation damage and has limited endurance (~10^6 program and erase cycles). To overcome these challenges several emerging non-volatile technologies are on the brink.

One of the promising technologies that has taken a lead is Ferroelectric Random-Access Memory (also called FeRAM), based on the charge being stored in the ferroelectric domains, which we shall be discussing in more details in Chapter 2 of this thesis. Another memory option currently under development is Magnetoresistive[3] Random Access Memory (MRAM) [19], which has the potential to provide high density, radiation hardness, and nondestructive readout (NDRO) operation with theoretically unlimited endurance. Major challenges for MRAM include reducing the drive currents, eliminating the cell instabilities due to magnetization and thermal imbalances, the difficulty of process integration as well finding niche applications so that they can be produced in mass to be cost effective. Ovonic Unified Memory (OUM) [19] (also called as *phase change memory* is based on a chalcogenide alloy which changes its phase: polycrystalline or amorphous and hence resistance after being heated.) is claimed to have higher densities and faster access than other NVMs. However, main issues here are integration challenges of new materials to the existing CMOS process as well as high write currents. OUM is still under development with its immature process. A good survey on the overall working principles and tradeoffs of the emerging NVM technologies appears in [18][21]. Key characteristics of these current and future nonvolatile memory technologies (the best ratings of the test chips are

[3] Employs magnetic storage devices based on the principal that the materials magnetoresistance [18] will change due to the presence of magnetic fields.

shown) are listed in Table I. (Endurance is defined as the maximum number of read and write cycles (for destructive memories); write cycles (for non-destructive memories) after which the memory cell looses data retention characteristics).

Table I Characteristics of Emerging Non-volatile Memories [19] [52] [53]

Characteristics	**Flash**	**FeRAM**	**MRAM**	**OUM**
Largest Array (Mb)	8192	64	16	4
Cell size (F^2)[4]	8-10	18-32	10-20	5-8
Endurance (Write Cycles)	10^6	10^{16}	10^{14}	10^{12}
Operating voltages (V)	2.7-3.6	1.37-3.3	1.8-2.5	1.8-3.3
Read/Write speeds (ns)	20/10000	40/40	50/50	50/50
Extra mask steps needed	6-8	2	4	3-4
Read type	ND[5]	D	ND	ND
Process (nm)	90	130	180	250

The major requirements for a non-volatile memory are fast read/write operations, radiation hardness, cost effectiveness and compatibility with currently used integrated circuit (IC) processing technology, high endurance and retention, and nondestructive readout capability. The present trend is towards single-chip solutions for cellular phones, pagers, smart cards, PDAs and similar small portable products. These devices are ideal for system-on-chip (SoC) solutions with reduced system cost and power along with increased performance. Such applications require the embedding of memory into the system. Increasingly most of the system is integrated onto a single chip. Some of the key characteristics of embedded memory are logic process compatibility, minimum added process cost, low voltage and low power operation, small cell size and a very large number of read and write operations before wear-out (i.e. high endurance). Ferroelectric

[4] F = feature size. The intended definition of feature is the minimum realizable process dimension but in fact equates to a dimension that is one-half the wordline (row) or bitline (column) pitch.

[5] ND = Non-destructive, D = Destructive

Random-Access Memory or FRAM™/FeRAM has attributes that make it attractive in non-volatile memory applications. It provides a reasonable choice for applications where sub-micro second programming is needed. FeRAM-based products have been available commercially for several years in limited quantities from Ramtron in the U.S and a few Japanese companies (e.g. Fujitsu and Matsushita) [41]. The technology is now emerging as a mainstream memory selection although it is only available through a limited number of foundries (e.g. Fujitsu).

1.2 Objectives

A major challenge currently barring the widespread acceptance of FeRAM is its modest density compared to Flash memory and battery backed up SRAM or DRAM. The cell size is reported to be as small as $18F^2$ [26]. Flash memory cells have been reported to be as small as $8F^2$ [19]. There have been various attempts to decrease FeRAM cell sizes and hence increase the chip density leading to Mbit capacities. The greatest reported density part is a 1T1C 64 Mbit test chip developed in 130-nm CMOS process by Texas Instruments [18]. One recent attempt to increase the density has been Chain FeRAM [7] a completely novel architecture design similar to NAND Flash memory.

The objective of this present work is to investigate a technique for increasing the density of FeRAM. The proposed technique is based on multilevel signaling in the memory cell. This concept is analogous to Multilevel DRAM (MLDRAM) and Multilevel Flash (ML Flash) [29]. Future generation DRAMs face the challenge of holding cell charges of a fixed minimum size (e.g. 50fC) in the cell of shrinking size. Increasing the cell capacitance can be achieved by increasing the dielectric constant of the cell capacitance dielectric. Since ferroelectric materials have a much higher dielectric constant and are already in use in FeRAMs, they have an excellent chance of being used as the dielectric material for future generation DRAMs. A very good survey on dielectric materials

for next generation Gbit DRAMs can be found in [23]. In such a case multilevel ferroelectric memories may be used to reach Multi-Gigabit (MGb) capacity.

The major objective in using multilevel memories is reduction in the cost per usable bit. The concept of multilevel storage was adopted successfully in Flash memories to increase the storage density with relatively minimal cost. This was due to a couple of reasons. The first and prominent one was reduction in fabrication costs. It is a very costly process to scale the devices since a huge amount of additional fabrication equipment is required [44]. Also, to implement a multilevel cell there is a very small amount of overhead in terms of area. And the fabrication lines do not require much change for implementing multilevel memories [44]. On similar lines, by means of multilevel cell technology, there is a potential to reduce the FeRAM memory cost. In this work we are trying to evaluate the practicality of such multilevel storage in FeRAM. The objectives being (1) defining the major challenges for such storage viz. designing adequate memory cells, improving noise margins, defining possible ways for read and write operations, visualizing the impact of parametric process variations and generating programmable reference voltage for reliable sensing and (2) proposing realizable solutions for these challenges. Three levels of storage on a single cell is used as a means to prove this concept and realize the objective.

1.3 Thesis Outline

This thesis assumes that the reader is aware of the basic structures and technologies in semiconductor memories like DRAM and SRAM. Background information on semiconductor memories in general can be found in [25]. The remainder of this thesis is organized into the following chapters: Chapter 2 reviews basic concepts in ferroelectricity and provides a literature survey on ferroelectric memory principles and some important circuit-level and architectural issues. Chapter 3 presents some of the concepts and design challenges in multilevel memory designs using examples in multilevel Flash and multilevel DRAM. Chapter 4 explains the novel idea of multilevel FeRAM design. Cell size

design issues and the read and write operations are discussed. Two techniques for the design of the required middle level voltage driver are evaluated. A reference voltage auto-calibration technique is proposed with some architectural proposals. Chapter 5 investigates the effects of statistical parameter variations and thus characterizes the robustness of the proposed design. A model based on the *central limit theorem* is used to study these variations. Finally Chapter 6 summarizes the work done in this thesis, discusses some of the advantages and disadvantages of the idea (feasibility analysis) and also proposes directions for future research.

Chapter 2

Background

2.1 Ferroelectricity

Ferroelectricity was discovered in 1921 [45]. Ferroelectricity can be described as a collection of quantum mechanical phenomena that create spontaneous dipole moments in the microscopic domains of materials leading to net electric polarization. Materials exhibiting this behavior are called *ferroelectrics* [1]. Despite the fact that none of these materials contain iron, the prefix "*ferro*" is used because they exhibit a hysteresis phenomenon similar to the remnant ferromagnetism in ferromagnetic materials. The interesting property of ferroelectrics is that the direction of the ferroelectric polarization can be switched by applying an external electric field. One of the better known ferroelectric materials is *barium titanate* ($BaTiO_3$). This material is but one member of a family of well-known ferroelectric materials called *perovskites* that share the chemical structure ABO_3. Another widely used material in this family is *lead zirconate titanate* (PZT) whose structure is described by the formula $Pb(Zr_xTi_1$.

$_xO_3$) [18]. Yet another family of ferroelectric materials is layered perovskite, which has oxide layers that are interleaved with perovskite layers in a lattice structure. One widely studied member in this family is *strontium bismuth tantalate* with formula $SrBi_2Ta_2O_9$ (SBT). Compared to PZT, SBT has superior endurance characteristics such as fatigue and imprint [1]. However, other properties (e.g. annealing temperature and remnant polarization) are superior in PZT. Most commercial FeRAMs appear to use PZT as the ferroelectric.

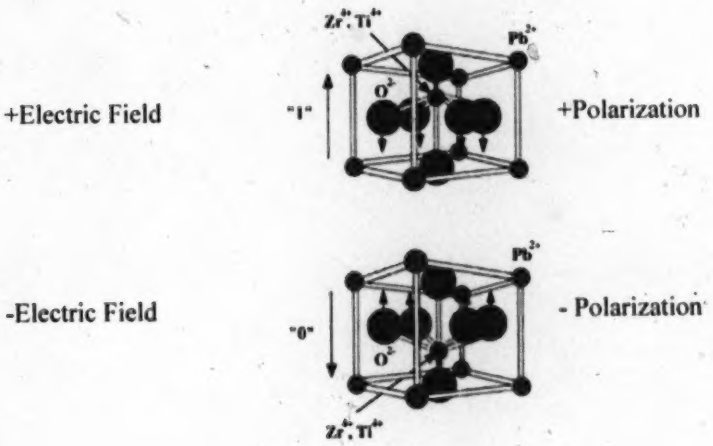
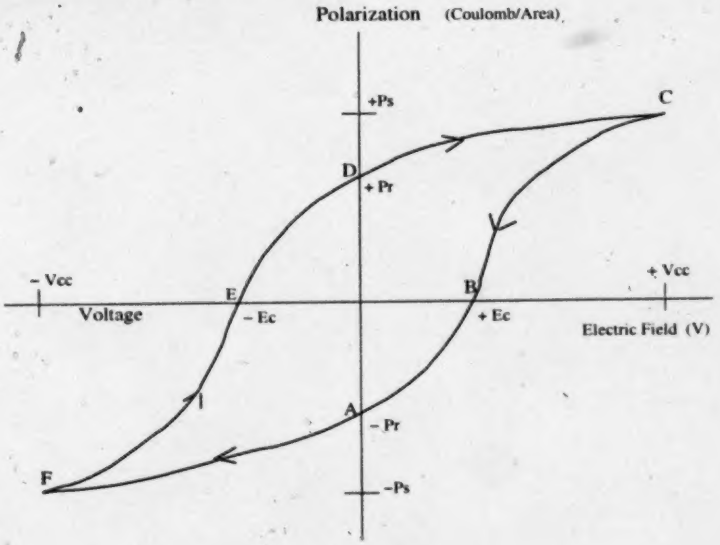


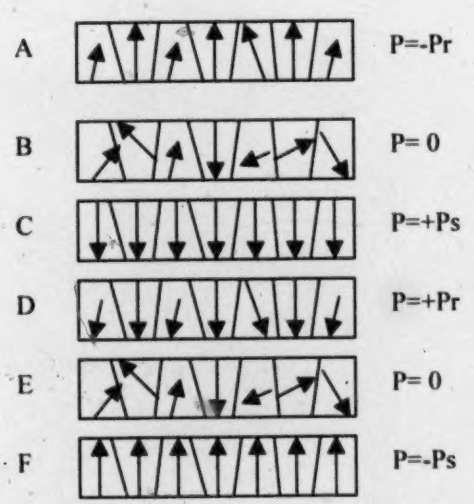
Figure 2-1 Dual-state Ferroelectric Lattice Under the Influence of Positively and Negatively Oriented Electric Fields [45]

Hysteresis describes one of the key characteristics of a ferroelectric material. Hysteresis behavior is evident in a plot of Polarization (P) versus Electric field (E). Ferroelectric hysteresis is best explained by noting the behavior of the microscopic domain dipoles in the ferroelectric crystal. In the absence of any external electric field, the microscopic dipoles are oriented randomly resulting in almost zero net dipole moment. Uniform alignment of electric dipoles is present in each domain in the crystal. In other domains of the crystal, the spontaneous

polarization may be in some other direction. The interface between two domains is called a domain wall, as shown in Figure 2-2, which is a simplified view of a thin film of dielectric material.



(a)



(b)

Figure 2-2 (a) Hysteresis Loop (Plot of Polarization vs. Electric Field) (b) Domain Changes in Response to Applied Electric Fields.

When an electric field is applied, the domains switch from their present direction of spontaneous alignment to another that is more closely aligned with the applied external electric field Thus ferroelectric material is a spontaneously polarized material with reversible polarization. As the applied field increases in strength, more dipoles get aligned in the direction of the electric field and hence the resultant macroscopic dipole moment is large and the material is said to be polarized. The degree of polarization is expressed as a charge per unit area. In a normal linear dielectric, when the electric field is removed, any polarization within the material reduces to zero; in the case of ferroelectric material this polarization is retained to some degree. This is due to inherent displacement of nano-scale polarizations due to atomic displacement within the crystal structure of the material. Figure 2-1 shows the cubic structure formed by Pb atoms: oxygen atoms are present on the square faces around the central Zr/Ti in each cubic cell. The applied electric field causes the central Zr/Ti atom to shift slightly in on direction while the O atoms shift slightly in the opposite direction. This changed structure is retained even after the field is removed. The two states of the structure can be used to represent the binary states "0" and "1".

2.2 The Ferroelectric Capacitor (FeCap)

A ferroelectric capacitor is formed by replacing the dielectric in a capacitor with a ferroelectric material. The relative dielectric constant[6] ϵ_r of a ferroelectric material is around 100-1000 [23], which is big enough to dump sufficient charge on the bitline from the switched[7] charge. The hysteresis loop can be observed when a sinusoidal external electric field is applied across this capacitor. As shown in Figure 2-2 (a) (with point B as the starting point), as the applied electric field E is increased the resulting polarization P increases roughly exponentially. However, after a certain point (point C in the figure) P only increases linearly, the

---

[6] In a parallel plate capacitor, the relative dielectric constant of the dielectric material is given by $C*d / (A.\epsilon_0)$, where C is the capacitance, d is the inter-plate distance, A is the plate area, and $\epsilon_0 = 8.85*10^{-12}$ F/m is the permittivity of vacuum.

[7] Refer to section 4 to understand the switching of charge during read operation

polarization at this point is said to be the *saturation polarization Ps*. The value of the polarization when E is then removed (i.e. E=0) (point D in figure) is called, the *remnant polarization Pr*. The charge corresponding to this remnant polarization is called the remnant charge Qr. This remnant charge is responsible for generating the voltage signal during sensing while reading and is non-volatile. On the hysteresis curve the magnitude for the reversed electric field that is just sufficient to decrease the net polarization to zero (point E) is called the *coercive field Ec*. At this point the remnant polarization reverses polarity if the reverse applied field is increased further in magnitude (moving towards point F). The arrow direction of increasing and decreasing polarization with increasing and decreasing applied electric field as shown in Figure 2-2 (a) is followed for all the hysteresis loops shown throughout this thesis.

2.3 Basic Memory Cell Structure

Integrated nonvolatile memory cells using ferroelectric films can generally be classified into two types of structures: (1) 1T-1C (or 2T-2C) and (2) MFSFET. Each cell structure has its advantages and disadvantages.

2.3.1 DRAM-like FeRAM cells

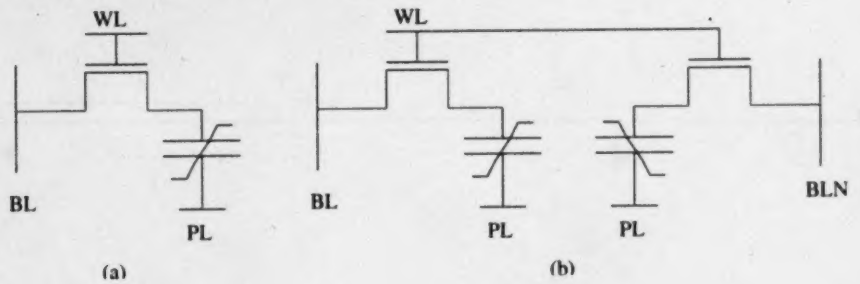


Figure 2-3 (a) 1T-1C cell (b) 2T-2C cell

The 1T1C FeRAM cell is similar to a 1T-1C DRAM cell and consists of a ferroelectric capacitor (C) to store data and a transistor (T) to allow access to the cell signal. Read operations rely on being able to sense charge shifts in the cell, which result from changes in polarization when a voltage pulse is applied to a

plateline that is connected to one terminal of the cell capacitor. The data signal stored as a dielectric polarization in a cell is destroyed in each read cycle; in other words, FeRAM has a destructive read operation. Therefore, the data in a FeRAM cell needs to be re-written after every read cycle. This slows down the reading operation compared to other memories (e.g. SRAM or Flash) that have non-destructive read operation. The present 1T-1C (or 2T-2C) architecture is a planar structure where one (or two) transistor(s) is (are) arranged with one (or two) capacitor(s) on a two-dimensional array. The 2T-2C structure has the disadvantage of having a larger cell area. [See Figure 2-4]. However, the 2T-2C uses a more robust self-referencing read operation that compares the complementary signals from the two cell capacitors.

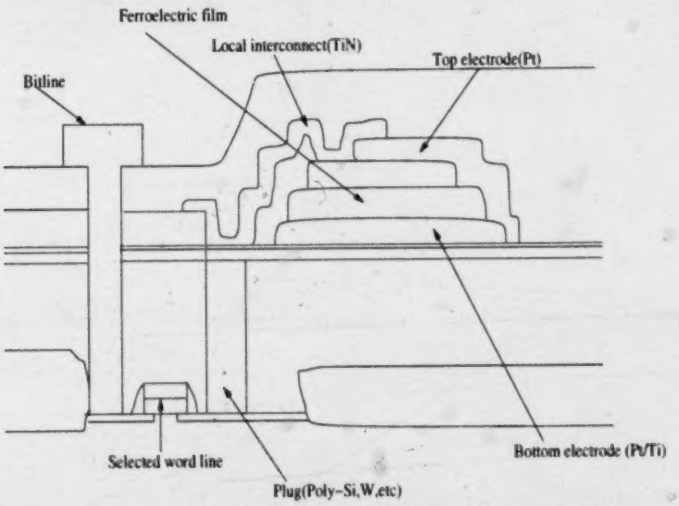


Figure 2-4 Ferroelectric Memory Cell Structure [24]

2.3.2 MFSFET FeRAM Cell

MFSFET (Metal Ferroelectric Semiconductor FET) is a field effect transistor that is realized by replacing the silicon dioxide film in a conventional MOSFET with a ferroelectric film under the gate [18]. The cell can be written by applying either a positive or negative voltage across the gate with both the drain and source terminals held at zero volts. Storage of a bit is based on changes in the transistor threshold voltage according to the polarization direction in the gate. Such a cell is

thus similar in operation to a cell in EEPROM or Flash memory; hence, similar circuit techniques can be used in a MFSFET cell. The MFSFET read operation is non-destructive and has the smallest cell area of all the FeRAM alternatives [24]. However, crystal inconsistencies at the interface between the silicon substrate and ferroelectric film make it difficult to control the threshold voltages of the transistor. [18]. Another issue is that, in the MFSFET, the ferroelectric film degrades rapidly to the extent that it cannot hold nonvolatile data. MFSFET has, therefore, not as yet been realized into any product [24].

2.4 Two-level Operation

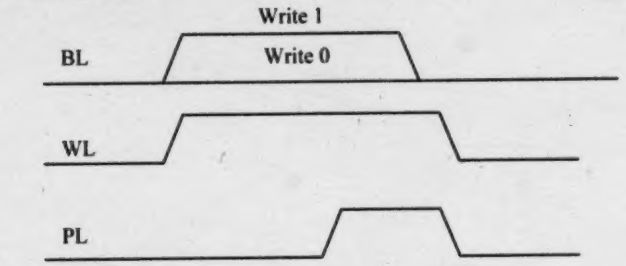
2.4.1 Read and Write Sequences

Write Operation

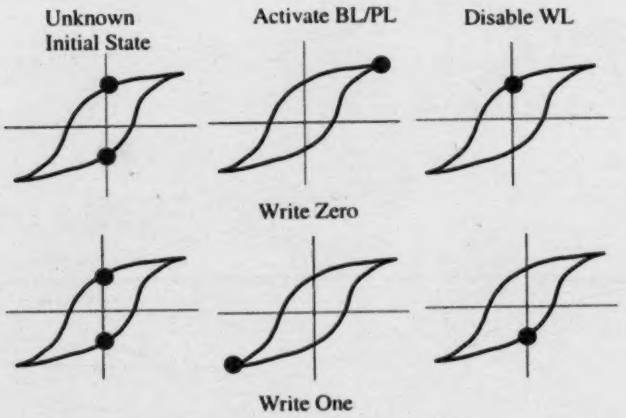
Figure 2-5 shows the write operation for a 1T-1C FeRAM cell structure. The bitline (BL) is driven to the voltage, either V_{DD} or V_{SS} that encodes the value to be written. Simultaneously the wordline (WL) is pulsed to turn on the access transistor of the addressed cell. Then the plateline (PL) is driven high. Since the PL capacitance is relatively large[8] (due to large number of cells connected to it and due to its own parasitic capacitance), the rise and fall times are very high (around 20 ns) depending upon the implementation technology [20]. Inside the material the resultant polarization (shown as a black dot for a given value of applied electric field) follows the sequence of operation as shown in Figure 2-5. The final polarization state (+Pr or –Pr) is retained upon deactivation of the WL. There are various ways described in the references concerning the sequence by which the WL, BL and PL are pulsed. In [1][6] the PL is pulsed as shown below while other schemes [3][18] hold BL=1, and PL=0 to write one and vice versa to write zero. However, as long as the effective voltage applied across the cell

[8] C_{PL} is of the order of 1-10 pF [20]. This depends on the technology of implementation, the number of cells connected to the PL and the capacitance of each cell.

capacitor remains positive to write one and negative to write zero, the two techniques are equivalent.



(a)



(b)

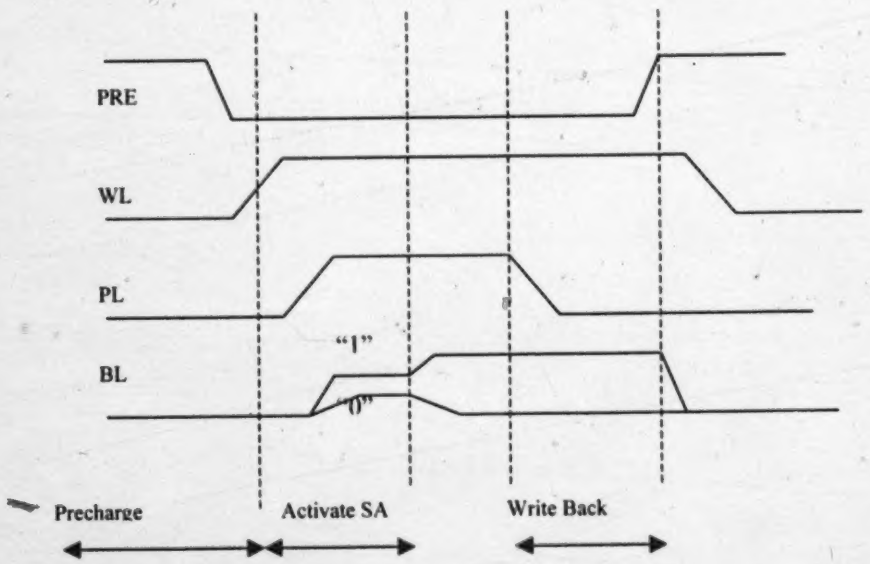
Figure 2-5 (a) Write Waveforms (b) Polarization States During FeRAM Write Operations

Read Operation

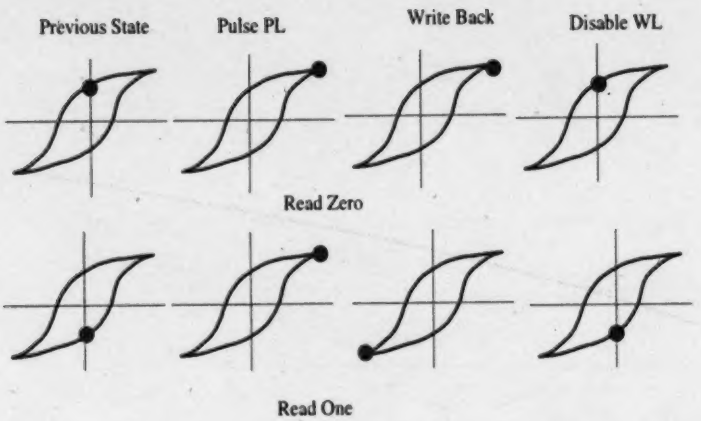
To read from a cell, the BL is precharged (PRE=V_{DD}) to '0' V and then isolated electrically (PRE=V_{SS}) before the WL is selected. Then the WL is pulsed (to V_{DD}), and V_{DD} is applied to PL so that, effectively, a negative voltage is applied across the ferroelectric capacitor. As shown in Figure 2-6(b), when the cell holds a '0', a relatively small quantity of charge is transferred onto the BL because there is no reversal of polarization[9] in the FeCap dielectric, consequently BL is bumped

[9] The cell is already negatively polarized.

up by only a small voltage V_L. On the other hand, when the cell holds a '1', there occurs a transfer of a larger quantity of charge because of the reversal (also called switching) in dielectric polarization, consequently BL is bumped up by a larger voltage V_H. These voltages are divided between the capacitive divider formed between C_{FE} and C_{BL}. A sense amplifier connected to BL with a reference voltage (V_{ref}) between V_H and V_L drives the BL to V_{DD} if the voltage on the BL is greater than some reference voltage V_{ref} and to zero if it is less than V_{ref}. It is important to understand here that the read operation is destructive because when '1' is read out, the data is destroyed and becomes '0' because of the reversal in the polarization of the cell dielectric. Accordingly, a '1' needs to be re-written to restore the original data. The voltage of the BL is V_{DD} after '1' is read out. This '1' can be re-written by setting the PL voltage to V_{SS} = 0V with the BL held at V_{DD}. The cell is thus restored with the original data of '1'. Reading out a '0' does not destroy the data because the polarization of the cell dielectric is not reversed. The data '0' is therefore still held in the cell after its state is sensed over the bitline. However restore still happens as the data is in the SA and WL is active.



(a)



(b)

Figure 2-6 (a) Read Waveforms (b) Polarization States During FeRAM Read Operations

The capacitance of the cell is denoted by C_0 or C_1 depending upon the stored charge. Here V_H and V_L are defined as follows:

$$V_L = \left(\frac{C_0}{C_0 + C_{BL}}\right) * V_{DD} \quad \text{when the stored data is '0'}$$

$$V_H = \left(\frac{C_1}{C_1 + C_{BL}}\right) * V_{DD} \quad \text{when the stored data is '1'} \qquad (1)$$

2.4.2 Sensing and Amplification

The sense amplifier (SA) plays a critical role in FeRAM because of the concerns about reliability issues (discussed in section 2.5.3) in ferroelectric memories. Reliability issues affect the sense amplifier operation by lowering the noise margins. The sense amplifiers can be classified as current sense amplifiers and voltage amplifiers. The most commonly used SA is the latch-type cross-coupled differential amplifier [25]. There are many factors that may affect the performance of the sense amplifier, such as transistor transconductance, threshold voltage, temperature, on resistance, channel width and length, etc [25]. A sense amplifier design is required that is tolerant to significant process variations and signal noise.

Two widely used sensing techniques are *step sensing* and *pulse sensing* [1]. In the former technique the SA is activated after the rising edge of the PL and almost after 90 per cent of the signal voltage is developed on the BL. This leads to faster sensing. On the contrary, in the pulse sensing approach the SA is activated after the whole PL pulse (both rising and falling edges) is applied. This approach is relatively slow; however, since the falling edge is applied, the non-switching part[10] (due to the linear cell capacitance) of the cell signal is eliminated from the signal seen by the sense amplifier. The non-switching part is subject to process variations and it is desirable to eliminate these variations from the weak cell signal. This becomes more significant with further scaling down in technology; in fact this has become evident in the latest fabricated 0.13 μm TI chip [16]. More details on these techniques can be found in [1][13].

Here we will consider some more issues related to sensing [3]. One challenge in the sensing operation is the effect of transistor and bitline noise imbalance during sensing. BL and BLN (the complement bitline) are the two differential inputs to the sense amplifier (SA). On one of these two inputs the stored charge from the cell is dumped (as well the precharge voltage to which the BL has been already precharged) while on the other a reference signal voltage is applied. However, the capacitance of the reference-carrying bitline is just C_{BL} while it is the sum of C_{BL} and C_{FE} in the bitline carrying cell signal. This creates a load imbalance during sensing. This causes the SA to drive one line slightly faster than other, hence degrading the sensing ability of the SA. One solution proposed in [3] is to de-assert the WL and thus disconnect C_{FE} from C_{BL} before enabling the SA, and later enabling the WL for a write-back operation. However, this introduces an access time penalty for all memory read operations. Another important issue is the precharging of the bitline. Conventionally in DRAMs the bitlines are precharged to $1/2V_{DD}$. In the case of FeRAMs the bitlines are all precharged to V_{SS}. This is due to the fact that the switching voltages are large enough to provide sufficient self-bias for the SA [3]. However, this switching

[10] Discussed in section 4.1

voltage is expected to reduce with smaller capacitor sizes as the technology scales down further (i.e. cell sizes smaller than 0.5 μm^2). Thus in the future it may be appropriate to precharge the BL to half V_{DD} or perhaps one-third V_{DD} [13] for stable SA operation.

The switching time of a ferroelectric capacitor was found to be around ~10 ns according to one recent statistical study [22]. In an experimental study done in [11] it was found that how much ever long the sense and re-write times are, the BL voltage of the worst cell did not exceed ~1700 mV. During sensing about 80% of the final readout voltage is developed in the first 20 ns. Hence it was concluded that a sensing time of about 20 ns should be sufficient for successful high speed FeRAM operation [22]. Innovations in ferroelectric thin film technology materials have enabled faster switching times for deep-submicron technologies. Switching speeds also depend on the implementation technology and the BL architecture.

2.5 FeRAM Design Challenges

2.5.1 Reference Voltage Generation

As mentioned in the previous sub-section, a reference voltage needs to be supplied to one of the inputs of the sense amplifier. This voltage has to be midway between the expected "0" and "1" readout signals, which will be denoted by V_L and V_H, respectively (as in [1]). One major hurdle in generating this reference voltage is that V_L and V_H themselves vary with process variations, temperature, supply voltage and material reliability issues like fatigue, imprint, etc. As a consequence, fixed-value reference voltage generators cannot be used on chip. Rather, a variable reference voltage is required to accurately track the process variations and any ferroelectric material degradation. In one of the two widely used techniques, one ferroelectric capacitor that is larger than the memory cell capacitors is employed per column [1]. This capacitor stores a zero so that the non-switching charge dumped onto the BL during a read operation is larger than

that dumped by a "0" from the memory cell, but is still smaller than a "1" from the same memory cell. This technique tracks the process variations and in general noise on chip. However, a challenge with this technique is that the reference capacitor gets accessed far more often than any data-storing memory cell and hence the reference cell fatigues (wears out) faster. This weakens the strength of the reference signal and the resulting shift in reference voltage can cause sensing errors. Also, determining the proper size of the reference capacitor is difficult. In a second reference generation technique [1], instead of a ferroelectric capacitor a dielectric capacitor is used to generate a fixed reference voltage level. The advantage here is that the dielectric is not subject to fatigue or other material variations. But again, designing a correctly sized capacitor and ensuring that it will maintain an appropriate voltage level throughout the life of the chip is a challenge. Note that process variations affecting the reference cells (with linear capacitors) will not necessarily track the variations affecting the data cells (with the FeCaps). Several other techniques using various sizes of the reference capacitors per column can be found in [1]. A recent paper [15] uses a current referencing technique in which a reference cell per row is used and the current is later mirrored throughout the columns to obtain a stable reference voltage. A reference generation technique should be comprehensive and robust enough to take variations in device parameters into consideration, as mentioned above, and should maintain the best possible noise margins between V_L and V_H.

2.5.2 Access Time Delay

FeRAM test chips have been designed with read access times as fast as 30-50 ns [16]; the read and write access times of commercially available FeRAMs are still around 100 ns [41]. The access times are mainly dominated by the large rise and fall times of the heavily capacitive PL [3]. As discussed previously, the relatively long rise time delays for the PL are due to both the relatively large parasitic capacitance and resistance[11] of the PL as well due to the large switching currents

[11] Since it is too difficult to process Platinium PL with higher thickness [11], the PL is thin leading to higher resistance.

of ferroelectric capacitors. The total capacitive load is contributed by one active row and a large number of inactive rows. In one of the architectures a rise time of approximately 25 ns is projected, leaving a very small margin for any improvements [3]. Hence determining an optimum PL architecture is very crucial in the development of FeRAM. In a conventionally segmented PL architecture, pass gates are used to isolate local PLs from the global PL to reduce the capacitance. A non-driven PL architecture is proposed and used in [11]. Here [see Figure 2-7] the PL is kept at $\frac{1}{2}V_{DD}$ constantly (non-pulsed) and the BL is driven to V_{DD} or V_{SS}. This achieves access times as low as 60 ns. However, there are several disadvantages of this method. The total differential voltage applied across the FeCap is roughly halved leading to weaker (given same dielectric material) polarization in the material. Most of the ferroelectric dielectrics require applied voltages of around 3V to produce full polarization making it necessary to apply such high voltages for full switching. Another problem in this design concerns leakage of the stored charge from the storage node SN. Since the storage node is floating while the WL is inactivated and there exists a reverse leakage path to the substrate through parasitic PN junction, the node leaks any free charge. Also, since the other plate of the capacitor is held at half-V_{DD}, the data is destroyed.

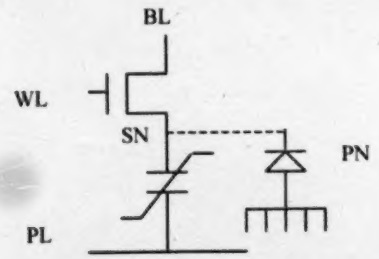


Figure 2-7 Non-driven PL Scheme

Hence this scheme requires a special refresh operation, which in turn increases the complexity of the design as well as the power. To overcome these problems a *bitline driven read scheme* [17] was proposed. In this scheme, prior to a read operation the bitline is precharged to V_{DD} while the PL is fixed at the ground potential and is not driven. When the WL is activated, depending on the ratio of

C_{BL} and C_{FE} (C_0 or C_1 depending on the previous cell state), a bitline voltage is developed. This voltage difference is sufficient for the sense amplifier to differentiate a "0" from a "1". This read scheme is 20% faster than the conventional scheme. A refresh operation is not required because PL is a 0V hence there is no destruction of the cell data [17].

More recently a new scheme has been proposed called the *differential capacitance read scheme* (DCRS) [13]. The authors claim to have eliminated the problems faced in previous schemes. As well this scheme increases the access speed remarkably achieving a decrease in the access time of around 40ns. In this scheme both the BL and BLN are initially precharged to ground. Then while keeping the PL at ground, the sense amplifiers are activated immediately after the WL goes high. The fundamental principle utilized in this scheme is that the voltage slew rate experienced by the BL when storing a "0" will be higher than when storing a "1". This is due to the fact that a ferroelectric capacitor storing a "1" has a lower capacitance than when storing a "0" [13]. This scheme achieves 40% reduction in read access time over the conventional read scheme as well as eliminating the issues encountered in both the non-driven PL and driven BL schemes. Commercial realization of this scheme is still under investigation. It is presently unknown whether just relying on the capacitive imbalances of the capacitors states ("0" and "1") is sufficient for long lifetime operations.

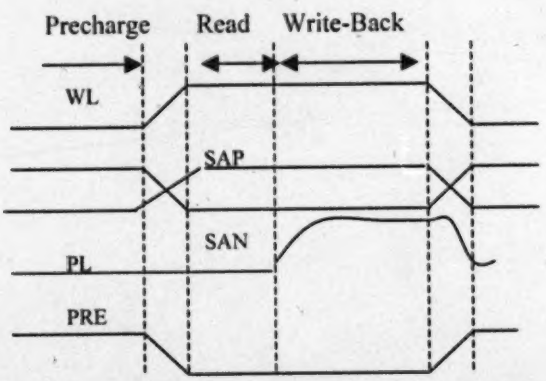


Figure 2-8 Read Operation in the Differential Capacitance Read Scheme [13]

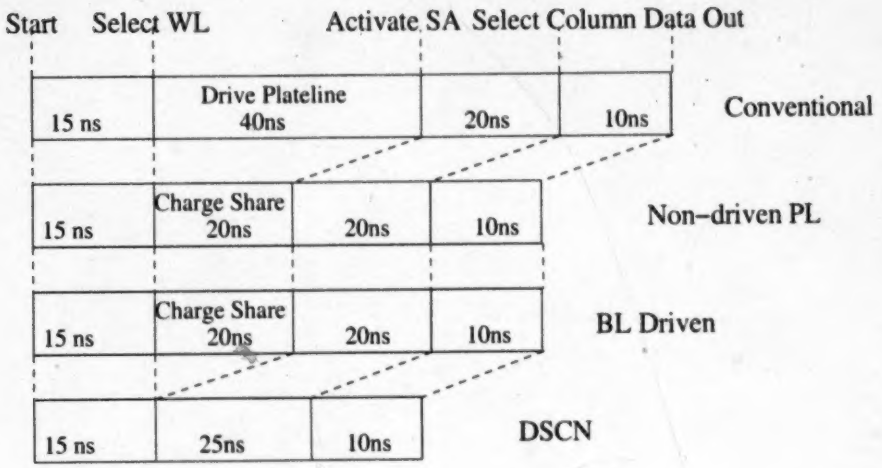


Figure 2-9 Comparison of Read Access Times [13]

A graphical comparison of the components of the read access time of these schemes is presented in Figure 2-9.

2.5.3 Reliability Issues

The characteristics of the ferroelectric material play a crucial role in the reliable operation of a ferroelectric memory cell. Reliability is especially critical in space and military applications or in any application where there are unfavorable environmental parameters [26]. There are challenges to be faced in materials design as well as in circuit design. For example, for ferroelectric thin films it is preferable to have a large remnant polarization so that a large polarization reversal current can be derived from a small-area capacitor. A low dielectric constant is preferred because a high dielectric constant material would produce a large displacement current that could interfere with the detection of the polarization reversal current. A low coercive field is preferred to permit low voltage operation of the FeRAM. Here we briefly discuss some of the reliability concerns that should be addressed when designing a robust ferroelectric memory cell. Refer to Figure 2-10 for accompanying illustrations.

Fatigue: Fatigue is a phenomenon in ferroelectric materials in which the remnant polarization (for the same coercive electric field) degrades with increasing number of polarization reversals (reads and writes) as shown in Figure 2-10(a). As an example, a minimum cell charge of one to two million electrons (160-320 fC) is necessary in a normal DRAM application to avoid soft errors [23]. From [3] a rough value of the saturation polarization in PZT is 30 $\mu C/cm^2$ and a rough cell area is $1\mu m^2$ implying a switched charge from the FeCap of roughly 600fC It has been shown that the level of available polarization in PZT moves below this limit after 10^{12} cycles. High temperature operation also increases fatigue [18].

Imprint: The tendency of a ferroelectric capacitor to prefer one state over the other if it stays in that state for a long period of time is called imprint [1]. Effectively this means that the FeRAM has become to an extent resistant to polarization reversal as shown in Figure 2-10(b). Imprint causes a shift (left or right) in the hysteresis loop, which reduces the signal margin. This becomes more critical as the coercive voltages[12] are scaled down with the scaling of the technology. Imprint characteristics are used to understand the data retention characteristic of FeCap.

Relaxation: This refers to the partial loss of remnant charge on a microsecond time-scale if the capacitor is left unaccessed after a sequence of cycling [1]. Also, polarizations return to their full value after data is rewritten. This effect is shown in Figure 2-10(c).

Data Retention: Due to aging, the polarization charge on the FeCap decreases hence reducing the sensing margins. Retention degradation depends on the type of material and environment in which the capacitor is used. Plate voltage bumps caused by accesses to other cells that share the same PL also cause data retention problems by disturbing the intended polarization of the FeCap.

Time-Dependent Dielectric Breakdown: Under constant voltage stress for longer periods of time the dielectric breaks down leading to memory failures [3]. The failure rate depends upon the characteristics of the material and operating

---

[12] This will shrink the height of the hysteresis loops (i.e. reducing the remnant polarizations).

temperature. This phenomenon must be considered in the design of DRAMs using ferroelectric capacitors.

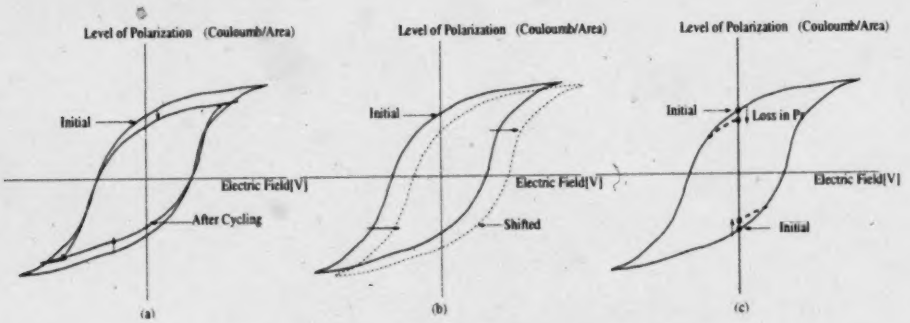


Figure 2-10 Effect of (a) Fatigue, (b) Imprint and (c) Relaxation

2.5.4 Memory Densities

Because of reliability issues, increasing memory densities and shrinking cell sizes are challenges in FeRAMs. This is because of the fact that addressing reliability issues requires higher noise margins, which are most easily achieved with larger cell sizes. Until roughly two years ago the cell structure was limited to 2T-2C. It was only after technology improvements in the process that lead to better endurance cycles, higher switching voltages and better material characteristics, that 1T-1C memories were considered reliable enough for 10 years of operation and started to become available for commercial use.

At present the highest density test chip that has been disclosed is 64 Mb [16] implemented in a 130-nm, five-level Cu/fluro-silicate glass (FSG) interconnect process technology with a 1T-1C cell architecture with access time of 30 ns. The test array operates at 1.37 V and consumes less than 0.8 mW/MHz. A density of 1.13 Mb/mm^2 is achieved with a cell size of 0.54 μm^2 and a capacitor size of 0.25 μm^2. If we express this in terms of the feature size, the cell size would be around ~32 F^2 where F denotes the minimum feature size.

Scaling trends in FeRAM need to be given equal attention since they seem to change significantly going from megabit to gigabit densities. Operating

voltages were scaling with $F^{1/3}$ until 0.25-μm technology but now they seem to be scaling more closely with F [3]. Bitline capacitance was scaling steadily with F in the megabit chips but for gigabit generations it seems to scale more like DRAMs [3] with $F^{2/3}$. This leaves a very limited margin in the C_{BL}/C_{FE} ratio indirectly implying more or less constant cell sizes for Gigabit memories. Getting smaller coercive voltages for ferroelectric thin films was a challenge over the last decade [3]. However recent experimental results for 60 nm and 85 nm have been shown to achieve coercive voltages at less than 1.2 V [3]. The switched charge Q_{sw} is proportional to the area of the capacitor. Though now the switched charge is scaling with F^2 it will have to be restricted (to maintain proper C_{BL}/C_{FE} ratio) since C_{BL} is not scaling with F^2 but rather with $F^{2/3}$.

2.6 FeRAM Architectures

Many architectural concepts in FeRAM have been derived from DRAM counterparts because of similarities in their structures as well due to the maturity of DRAM's architectural concepts. The common objectives are to get the minimum cell area and the maximum number of cells per chip, with the least power. There are several interesting architectures proposed for FeRAMs, each designed specifically to solve some issues suitable to the design requirements and to improve performance. The folded bitline architecture is a very common design feature that reduces the common mode noise and improves the noise margins [25]. Depending upon the design requirements appropriate orientations of WLs, BLs and PLs are chosen. For example in the WL||PL (WLs in parallel with the PLs) architecture [See Figure 2-11(a)] the entire row has to be accessed but this simplifies the routing of the PL [1]. In the BL||PL architecture [See Figure 2-11(b)] a single cell can be selected by simultaneous selection of WL and the PL. The need for a Y-decoder is eliminated for the PL. The SA is activated by the same PL signal and only one SA needs to be activated at a time thus reducing the power. However, activating the PL can disturb the cells in the same column just

as the cells can be disturbed for the entire row in the WL||PL architecture with PL shared between adjacent rows.

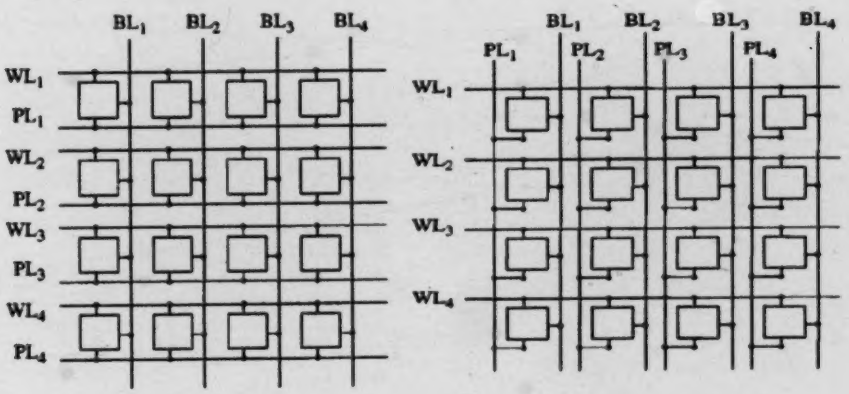


Figure 2-11 The (a) WL||PL and (b) BL||PL architectures [1].

The merged WL/PL architecture [See Figure 2-12] is similar to the 2T-2C architecture with opposite polarities written into C1 and C2. Here ML1 and ML2 act as the WL and PL. A read operation is performed by simultaneously pulsing ML1 and ML2. Later the sensed data is written back into the cell. The merged WL/PL architecture is said to achieve higher densities due to reduced number of access wires, which in turn is due to the merging of WL and PL [1]. However, there is a penalty of more complicated processing needed to stack the capacitor terminal onto the transistor gates as well providing side contact to source or drain.

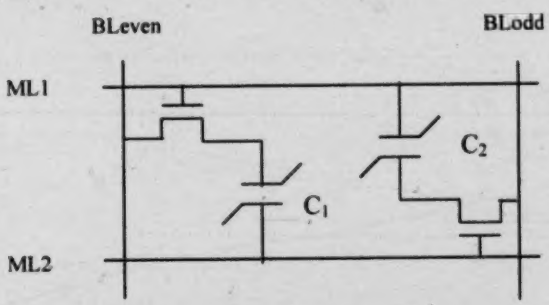


Figure 2-12 Merged Plateline Architecture [1].

Shadow SRAM [18] is a dual-mode architecture in which the ferroelectric capacitors are only used during power up and power down. Right after power up data is read from the FeRAM cells by pulsing the PL. Once the data is read the FeCaps are isolated from the operating SRAM. Data is restored back once again to the FeRAM before power is withdrawn from the SRAM. This helps in reducing the amount of fatigue on the FeRAM cells and increases effective endurance of the SRAM-FeRAM combination.

Figure 2-13 below shows a 1T-2C cell, called a transpolarizer in [1], in which two ferroelectric capacitors are connected to the source of one access transistor. The capacitors have separate platelines shown as PL1 and PL2 in Figure 2-13. A "0" is stored by writing a positive voltage signal in both capacitors and a "1" by writing a negative voltage signal. This is achieved by pulsing both the PLs and writing the required voltage onto the BL. In a read operation, BL is precharged to $V_{DD}/2$ with only PL1 being pulsed. Write-back is achieved after sensing and pulsing both the platelines. The main advantages of this architecture are that the half-V_{DD} reference voltage is unaffected by FeCap degradations and the capacitors needed are quite small. However, two platelines are needed.

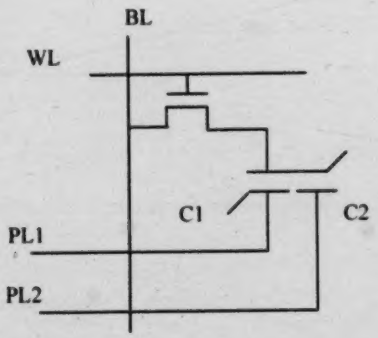


Figure 2-13 1T-2C Cell Structure

There are several other architectures, which are specifically designed to reduce the plateline delay and increase the access time including non-driven PL,

segmented PL, bitline-driven etc. These architectures are discussed in section 2.5.2. One design architecture worth discussing here is *chain FeRAM* [7]. As shown in Figure 2-14, it consists of memory blocks in which memory cells are connected in series as chains of cells. One end of the cell block (chain) is connected to cell PL and the other is connected to BL via a block-selecting transistor. The beauty of this design is that a cell size of as small as 4 F^2 is achieved using only planar transistors. As in a NAND Flash memory, the cell area can be reduced by sharing contacts among adjacent cells. At a steady stable state both the nodes of the FeCap are short circuited by the turned on cell transistor. As a consequence, both the non-driven and half-Vdd, as well as driven cell-plate scheme are applicable.

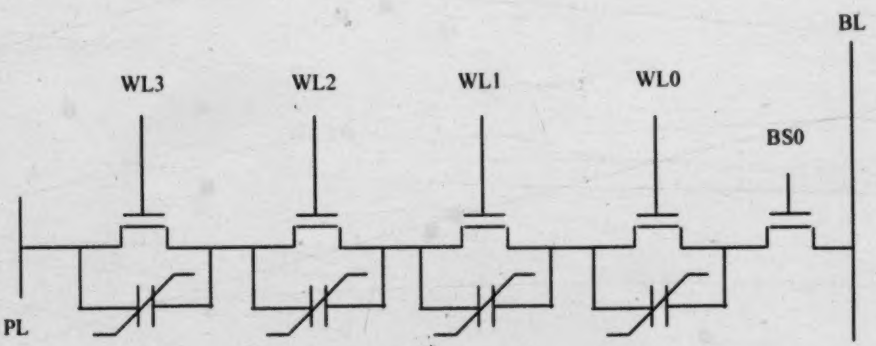


Figure 2-14 Chain FeRAM

During the read operation the block-selecting transistor is on while the access transistor is turned off. The BL is precharged to V_{SS}, hence a voltage of half-Vdd is applied between the BL and the PL and the cell data is read out. However, there are certain issues to be taken into account before using this design technique. In terms of reliability, the readout voltage in a read operation in a block of cells shows variations due to voltage drops in the path[13] transistors. This readout voltage seems to change with the chain length, i.e. the number of transistors in a block. Second, since all the access transistors in the cell are always on with boosted

[13] Unselected transistors, which come in the path of BL, selected transistor and PL.

voltage applied across them, their gate oxide degrades more easily and generates reliability concerns over the lifetime of the cell. Third, since the transistors in a block are in series, the plateline capacitance increases resulting in more delays in the plateline. A more detailed review of these architectures can be found in [1][3][18].

Chapter 3

Review of Multilevel Memory Technologies

3.1 Overview

The most important driving factor in memory development is to increase the density of data storage, while maintaining or reducing the cost per bit. The conventional way of achieving this goal in semiconductor memories is to reduce the physical dimensions of the memory cells. For this designers have to solely rely on the geometry scaling achieved with the help of advances in semiconductor processing. Advances in DRAM cell design allow the footprint of each cell to be scaled down while maintaining a minimum soft capacitance of at least 30 fF (a lower limit that is determined by the greatest acceptable soft error rate in the present environmental radiation). Another possible direction for increasing the storage density that uses circuit-level techniques is multilevel storage [27][30][31][32].

The main idea of multilevel storage is to increase beyond one bit, the amount of data being stored on a single cell and hence to reduce the cost per bit. There have been multilevel implementations of nonvolatile Flash as well as of

DRAMs. However, only multilevel Flash has become commercially available [27]. Multilevel DRAM implementations are still under research. Multilevel DRAM has evidently not yet proved to be cost-effective with respect to the various trade-offs involving yield, reliability, complexity and compelling applications. The unfavorable conditions for multilevel DRAM may change, however, with the availability of improved dielectric materials and rapidly rising lithography costs.

3.2 Multilevel Flash Memory

Flash memory has been in production for 20 years and is commercially available in market with densities as high as 8 Gbit[14]. Multilevel Flash (ML Flash) stores varying amounts of charge on the floating gates (FG) of the access transistor to the Flash cells. The charge on the floating gate shifts the turn-on threshold voltage of the access transistor. These voltage shifts are converted or mapped to digital data with a form of A/D conversion at the sense amplifiers.

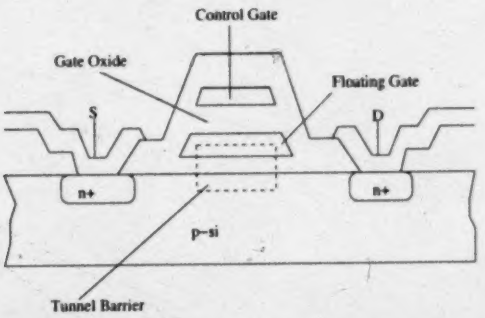


Figure 3-1 Floating Gate Memory (Flash)

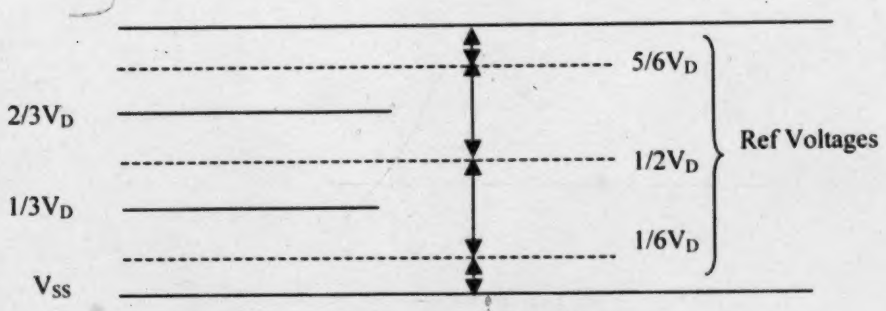
ML Flash memory design faces challenges at the architectural level as well as at the circuit and device physics levels. These challenges are primarily due to the challenges of: writing precise amounts of charge on the FG (programming), reliably sensing/ reading the weak cell signals on the bitlines and the long term

[14] This is a Nand Flash available from Samsung

retention of the stored charge on the FG so that the varying signal levels do not merge and become indistinguishable.

3.2.1 Programming Techniques in ML Flash

In ML designs the programming needs to be accurate to properly set the V_T of the FG transistors. The V_T shift is controlled by accurate charge transfer (I_{OX}) to the FG, which in turn depends on the proper biasing conditions. Two widely used



Levels	Binary Data	Reference
V_{DD}	"11"	
$2/3V_{DD}$	"10"	$5/6V_{DD}$
$1/3V_{DD}$	"01"	$1/2V_{DD}$
V_{SS}	"00"	$1/6V_{DD}$

Figure 3-2 Four-level Signaling (2 bits per cell)

techniques in bi-level memories are: *Carrier Hot Electron Injection (CHE) and Fowler-Nordheim Tunneling (FNT)* [25]: CHE happens when the electrons in the channel of the MOSFET gain a sufficient amount of energy while in transit such

that they get injected on to the floating gate because of a horizontal electric field applied between the controlled gate and the channel. In FNT electrons tunnel into the FG due to the applied gate-to-substrate high electric field. Compared with FNT, CHE requires lower voltages and achieves high speed, better endurance, less circuit overhead with weaker disturbances in the array [29]. On the other hand, FNT requires much lower power consumption thus allowing larger programming parallelism, and increasing overall throughput. FNT can be used for both writing and erasing; however, CHE cannot be used for erasing because it is not possible to heat electrons in the floating gate by means of an external electric field. For fixed bias conditions there exists a linear relationship between the applied gate voltage V_G and ΔV_T [29]. However, due to process variations the distribution of V_T (V_T^1, V_T^2 V_T^3, etc.) is not narrow and adjustments in the programming voltages have to be made. In the following some of the common programming techniques are briefly discussed.

P-V (Program and Verify) techniques: These techniques are self-convergent programming mechanisms providing a well-defined final state.

Gate-Voltage Programming: This uses the CHE injection method, which exploits the linear relationship between ΔV_T and V_G. A staircase waveform for the gate voltage is increased at each program step by a fixed amount of say ΔV_G. Suitably narrow V_T distributions are thereby achieved.

Drain Voltage Programming: In this method V_G is kept constant while the drain voltage V_D is varied. Memory cells belonging to selected word lines can be programmed to different levels in the same programming period by providing the required different V_D voltages.

Self-Controlled Programming: P-V techniques are iterative and need more time than single-step methods. To reduce this time the cell V_T can be controlled while it is being programmed [29].

Drain Current Monitoring: While employing CHE it has been observed that a one-to-one relationship exists between V_T and drain current I_D for constant V_G and V_D [29]. This current is monitored to find when the required V_T is achieved. With accurate design and iterative tries acceptable precision can be achieved.

Self-Controlled FN Programming: As the name suggests, with the help of FN tunneling this technique stops programming automatically as soon as the target value of V_T is reached. This technique exploits the exponential dependence of I_{OX} and F_{OX} (oxide electric field). Programming is stopped simply by slightly decreasing F_{OX} by employing suitable feedback circuitry or embedded control mechanisms.

Off Chip Programming: This technique is used to program the cells with higher number of levels, thus eliminating any on chip programming circuitry and their issues. The technique achieves very narrow V_T distribution [29].

3.2.2 ML Sensing Schemes

As mentioned before, the sensing method plays a very critical role in the design of ML memories. In fact, the number of levels that can be reliably stored on a single cell is limited by the sensing margins available with the sense amplifiers. Various parameters need to be looked at such as sensing time, power, complexity and area. Compared to conventional bi-level memories, ML memories have more stringent requirements that become more severe with increasing numbers of signaling levels.

Basically, sensing the contents of an ML cell involves A/D conversion from 2^n levels to n bits. Both current sensing as well as voltage sensing has been used [29]. In the latter approach, the current drives a constant load to generate a voltage, which is later compared with the reference voltage. As far as sensing architectures consisting of sense amplifiers are concerned, either *parallel* or *serial* sensing can be used. Parallel sensing is nothing but a flash-type A/D converter, which compares the read current with $2^n - 1$ reference currents. This ensures fast conversion, i.e. high speed sensing. However, the disadvantages of flash conversion include more area and power consumption, which increase roughly exponentially with n.

Two known serial sensing methods are sequential serial and dichotomic serial sensing. In the sequential serial technique, different gate voltages are applied successively to the selected word line [29]. The cell current is successively compared with increasing reference currents starting from the lowest. Sensing is stopped when the reference current becomes larger than the cell current. This sensing technique is inherently slow due to the settling time requirements for the varying read voltage. In dichotomic serial sensing the reference is varied in a divide-by-two fashion [29]. The reference range is divided into two sub-ranges and comparison is started from the voltage, which is nearer to the middle of the range. Only one SA is needed in this approach with a successive approximation register, which stores the result of each comparison and controls the selection of reference current used at each search step. This reduces the number of sensing steps to *n*, i.e. to the number of stored bits per cell. This technique is most acceptable for the page-mode access patterns used in DRAMs since the increase in random sensing time does not affect the average read throughput for a burst of data from the same page.

3.2.3 Reliability Issues

For any technology, reliability issues come into the picture when considering long-term usage of the device. In the case of ML Flash there are two critical parametric limitations that decide the bounds. The first requirement is the V_T spread (or window). It is important to keep adequate spacing among the stored levels so that reliable sensing is possible throughout the life of the cell. The second conflicting requirement is to minimize the spacing between the signal levels so that the device operates for a limited spread of V_T. This is because an increased spread leads to larger charge transfer through the oxide and, in practice, to higher programming voltages, thereby worsening problems related to charge trapping within the oxide and excessive oxide leakage (degrading data retention) [29]. The endurance limitation of a properly designed ML Flash memory cell is

roughly the same as in a conventional bi-level Flash memory cell, i.e. 10^6 program /erase cycles. Some of the potential reliability issues in ML Flash include the following:

Erratic bit behavior: It has been found that a few bits per million sometimes exhibit a sudden change of tunneling current from a normal or very large unknown value causing widely different V_T shifts [29]. This effect has been attributed to trapping and de-trapping of holes in the oxide during tunnel conduction.

Data retention: This is the most important reliability issue for any ML memory. As mentioned in [29], accelerated tests have shown a maximum V_T shift of about 0.1 V within a device lifetime. This was found to be compatible for ML storage for 3-4 bits per cell.

Read Disturb: This is due to tunneling injection of electrons into the FG of a cell. This occurs during reading the cell or a nearby cell in the same word-line. This affects mainly the memory cells that are in the lowest V_T state. ML cells are more sensitive to read disturbs because the read gate voltage is higher. Read disturbs are more probable when the effect of cycling are considered. This is due to the oxide degradation caused by the high fields used during write operations [29].

Stress induced leakage current (SILC): Even at low electric fields a high-field stress on a thin oxide increases the current density. This additional current causes a significant deviation of the I-V curve from the theoretical Fowler-Nordheim characteristic at low field and is called SILC. In short it is a stress-induced oxide defect. SILC affects data retention and read disturb is more pronounced. SILC is mainly dependent on the oxide thickness.

3.3 Multilevel DRAM

The concept of MLDRAM evolved in the late 1980s in an effort to increase the density of memory chips and, in turn, reduce the cost-per-bit in the cut-throat competition of the DRAM market. Although DRAM feature sizes have scaled down greatly since then, the actual cell area expressed in terms of the minimum

feature size has remained the same ($8F^2$). The most recent efforts to reduce the actual cell area have been through using stacked or trench capacitors [29]. Using such complex cell designs the minimum cell area has remained $8F^2$, where F is the minimum feature size. Hence a new technology, which could further reduce the cell size, would be greatly desirable. Multilevel storage increases the per-cell storage density by storing more than one bit in a single cell. The concept parallels that of ML Flash discussed in Section 3.2.

3.3.1 Writing Schemes

As we saw in ML Flash, the writing accuracy is very crucial for multilevel storage because of the narrower noise margins. Depending on the architecture, various writing schemes have been proposed. In [28] four bits are stored in a cell. The data to be written (out of the 16 levels) is buffered in a special register in the column. A descending staircase voltage waveform is applied to the addressed WL while bitlines are set to V_{SS}. During the n^{th} step of the descending staircase voltage, because of the data value corresponding to n^{th} step stored in column register, BL starts to rise from V_{SS} to V_{DD}. Thus one of the available sixteen levels is written by raising the bitline during the appropriate step period of the descending WL signal. In Furuyama's design [30] the bitline is split into three equal sections called *sub-bitlines*. Charge sharing is used to form the required level with a combination of V_{DD}, $2/3V_{DD}$, $1/3V_{DD}$ and V_{SS}. Each section is written with a combination of V_{DD} and V_{SS}. For example if V_{DD}, V_{SS} and V_{SS} are written to three sub-bitlines at equal capacitance, we will get an equalized voltage of $1/3V_{DD}$ when we connect them together during charge sharing.

3.3.2 Sensing and Reference Generation Schemes

The sensing techniques proposed in the MLDRAM literature are similar to those that have been used in ML Flash. Parallel sensing, sequential sensing and a mixture of the two are the main approaches. Reference generation, being an important topic for MLDRAMs, is discussed in more detail here. The various techniques used in previous designs are now reviewed.

Global reference: In this technique, the reference voltages are generated outside the memory core [30]. Dummy cells within the data cell array are used to store these reference values. The reference voltages thus generated externally must be distributed through the MLDRAM circuitry. The advantage of this method is that it is simpler to generate such voltages as the generator is out of the core array. A major challenge is the design of a noise-free distribution network, which might create offset errors in the delivered references. Such global references could be generated on-chip or off-chip.

Locally generated reference: In this method the reference voltage is generated locally, i.e. within the MLDRAM cell array. References are generated by using appropriate capacitance ratios of the sub-bitlines. In Gillingham's MLDRAM [31], cell columns are formed using two pairs of sub-bitlines. The sensing employed is a two-step successive approximation. The references are generated sequentially depending on the result of the previous sensing (MSB to LSB) [31]. For the initial MSB sensing operation, one copy of the cell data is compared with ½ V_{DD}. Depending on the result of this comparison, the second comparison is done with 1/6 V_{DD} or 5/6 V_{DD} as in dichotomic sensing discussed in the case of ML Flash. However, these two references are generated from V_{DD} or V_{SS}, depending on the sensed MSB. Sensing the LSB begins by writing the sensed MSB value into the original data cell, and then out of the three sub-bitlines used in the design, two carry the MSB value while the third one is precharged to ½ V_{DD}. A switch matrix is used to connect these three sub-bitlines and generate the required averaged reference voltage of either 1/6V_{DD} or 5/6V_{DD}. A generic equation for charge sharing as provided in [32] is

$$V_{signal} = (V_{cell} - V_{cell-plate})(C_{cell}/(C_{bitline} + C_{cell})) + V_{bitline} \quad (2)$$

The advantages of Gillingham's method compared to Furuyama's method are that there is less degradation in the signal quality and less sensitivity to process variations. The main disadvantage is the increase in the sensing time. Also the bitlines need to be balanced for sensing; this is done by using dummy cells in the design.

Another method for generating the required reference, which is similar to charge sharing involving bitlines, is capacitor coupling. In this method required reference voltages are generated by charge sharing the sub-bitlines with the built-in-capacitors. However this method is more sensitive to process skews due to mismatches in the built-in-capacitors. Furthermore these capacitors cause imbalances during sensing. A detailed discussion of these MLDRAM techniques can be found in [32].

3.4 MLDRAM Challenges

As we have seen in the case of ML Flash, MLDRAM also has to overcome several stringent requirements before successful commercial realization. The main challenge, however, remains managing the narrow noise margins between adjacent signal voltage levels. This is entirely limited by the sensing margins available from the sense amplifier design. Though the operating voltages have been scaling down with the technology, thus limiting the noise margins between the levels, the sensing margins (minimum sensing signals) of sense amplifiers[15] haven't improved over the years. As the number of levels in a multilevel design increases, physical device imperfections increase more, causing increased cell leakages, soft errors due to alpha particle radiation, charge injection effects caused by wordline switching, reduced retention times etc.

---

[15] This is for the simplest latch-type cross-coupled sense amplifier

Input offset issues in sense amplifiers also lead to unreliable sensing of the cells. More advanced sensing amplifiers, like *charge-and-split* SA [48], *differential sense amplifier* without positive feedback and *Euclidean-Distance* sensing [49] might be required as has been used in ML Flash with the requirement of being small enough to fit in the specified pitch for high density memories. There are also issues, which need to be tackled at the architectural level to fit the economics of the DRAMs.

Another important issue in ML sensing is longer sensing time and this time increases exponentially as the number of levels increases. Assuming that there will always be a certain number of cells, which are leaky or have tendency to be faulty and their failure cannot be avoided, error correction techniques might be employed to achieve acceptable reliability. Note that these error correction techniques add overhead in area, power and timing. But they also relax the raw cell yield constraints, thereby leading to a possibly faster time to market.

3.5 Summary

In this chapter the concept of multilevel storage was reviewed. We saw how storing more than one bit in a single cell can increase the per-area storage density. Previous work on DRAMs and Flash memory cells was reviewed. Programming, sensing and reference generation techniques were discussed for each of them. Finally, we discussed some of the major challenges in MLDRAMs. MLDRAM has not yet become a commercial reality because of the already low cost-per-bit of conventional DRAM in the hyper-competitive commodity memory markets. However, four-level signaling was successful in the Flash business because of the higher amount of cost of two-level Flash, thus improving the cost per bit. The same analogy might apply to Multilevel FeRAM compared to two-level FeRAM.

Chapter 4

Multilevel Ferroelectric Memory Design

The concept of multi-level storage in ferroelectric memories appears to be a completely new idea. The multilevel concepts are similar to those in multilevel DRAMs. However there are new circuit-level implementation challenges. There is an increased need to accurately characterize the shape of the hysteresis curve so as to be able to depolarize the ferroelectric material to a zero or near-zero polarized state. In a 3-level FeRAM there are three possible states for the cell: one with the ferroelectric material positively polarized, one with it negatively polarized, and one with it depolarized [see Figure 4-1]. Also, it must be possible to transition from each of the three states to all of the other states, with same sequence of control operations, so that there is no time difference for reading and writing different states. All cells in the same row should see the same wordline and plateline signals. Only the column signals can differ (e.g. be data-dependent) going along a row of cells.

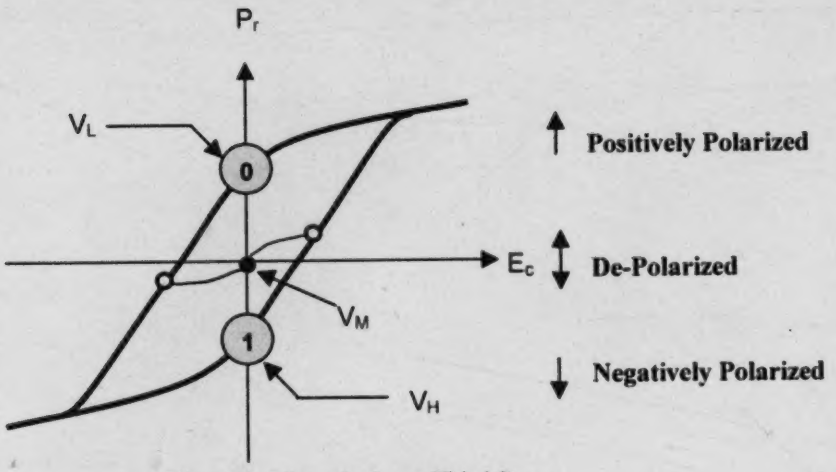


Figure 4-1 Introducing a Third State

4.1 Multilevel Cell Design

Determining an adequate cell capacitor size for a given bitline capacitance and number of cells in the array is critical in FeRAMs. It has been argued [3][13][14] that merely increasing the cell capacitor size does not always achieve greater bitline voltages. In FeRAMs, in fact, it is most important to determine the ratio of C_{BL}/C_{FE}[16] as described in [3] and [15]. The readout voltages (the BL voltages before sensing) are very much dependent, as it will be evident soon, on this ratio. Since C_{BL} depends on the array size and all the devices connected to the BL, it is essential to accurately estimate the value of C_{BL} before fabrication. The relationship between C_{BL} and C_{FE} arises because the cell charge, which encodes the stored data, is shared between the capacitor divider formed between them. To ensure that the cell dielectric is driven to full saturation, the voltage across the cell capacitor after the capacitor has dumped its polarization (switched) charge ($2Q_r$) onto the bitline should be roughly twice as large as its coercive voltage V_c [3]. This imposes a lower limit on C_{BL} as expressed by:

---

[16] C_{BL} is the parasitic bitline capacitance while C_{FE} is the ferroelectric cell capacitance corresponding to the dielectric capacitance (C_{ns}) and does not include the remnant charge Q_r.

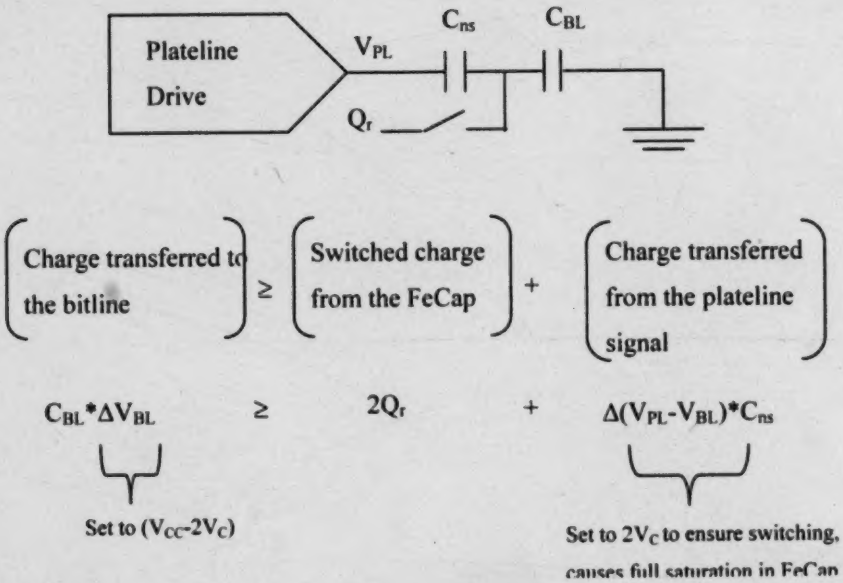

Plateline
Drive
V_{PL}
C_{ns}
C_{BL}
Q_r
(Charge transferred to the bitline) ≥ (Switched charge from the FeCap) + (Charge transferred from the plateline signal)
$C_{BL}*\Delta V_{BL}$ ≥ $2Q_r$ + $\Delta(V_{PL}-V_{BL})*C_{ns}$
Set to $(V_{CC}-2V_C)$
Set to $2V_C$ to ensure switching, causes full saturation in FeCap

Figure 4-2 Charge Transferred to BL due to Switching and Non-switching Charge

$$C_{BL} \geq \frac{2Q_r + 2V_c * C_{ns}}{V_{cc} - 2V_c} \qquad (3)$$

C_{BL} = Bitline Capacitance, Q_r = Charge corresponding to remnant polarization

V_c = Coercive voltage, C_{ns} = non-switching capacitance

At the same time, the bitline capacitance has to be small enough to allow enough signal voltage to be established on the bitline. This signal voltage limit will depend on the minimum reliably-resolvable sense amplifier input signal sensing strength; this is typically about 70 mV (V_{SE}) in the case of conventional FeRAM [3]. The Thevenin equivalent circuit with respect to the switched charge Q_r is:

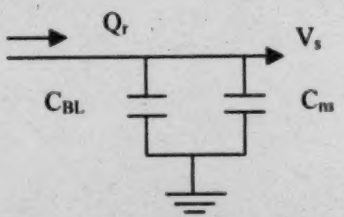


Figure 4-3 Thevenin Equivalent Circuit for the Switched Charge

If the switched charge Q_r is dumped into the circuit node, the change in voltage is derived from $V_s \{C_{BL} + C_{ns}\} = Q_r$, where V_s is the available signal on the bitline. Thus $C_{BL} = Q_r/Vs - C_{ns}$. The cell design determines Q_r and C_{ns}. To keep Vs above some minimum value V_{SE} we will have the following inequality. $Vs > V_{SE}$

$$C_{BL} \leq \frac{Q_r}{V_{SE}} - C_{ns} \tag{4}$$

Note, only one Q_r is used because the signal is with respect to a reference signal in between the "1" and "0" signals. The capacitance of a ferroelectric capacitor arises from two effects, namely, *non-switching polarization* and *switching polarization* leading to two types of capacitance. Switching capacitance (C_s) is caused by small shifts in atomic positions within the lattice. Non-switching capacitance (C_{ns}) is due to the movement of conduction electrons (and holes) in nearby conductors and semiconductors.

$$C_{FE} = C_s + C_{ns}$$

$$C_s \equiv \langle \epsilon_0 * \epsilon_r * A \rangle \div d$$

$$C_{ns} \equiv \langle P_{ns} * A \rangle \div Vcc$$

$$Q_r = P_r * A \tag{5}$$

Here ϵ_0 is the permittivity of free space and ϵ_r is the relative permittivity of the ferroelectric material, A and d are the area and distance between the capacitor plates, P_{ns} is the non-switching polarization and V_{cc} is the positive voltage supply. Q_r and P_r correspond to remnant charge and polarization, respectively.

Hence equations (3) and (4) imply the following two inequalities:

$$C_{BL} \geq \frac{V_c * P_{ns}}{(V_{cc} - V_c) * V_{cc}} * A \tag{6}$$

$$C_{BL} \leq \left(\frac{P_r}{V_{SE}} - \frac{P_{ns}}{V_{CC}} \right) * A \tag{7}$$

In the case of a cell used to store three levels, in equation (3) we should replace $2Q_r$ with Q_r and V_{cc} with V_M (the middle level voltage to be written) since we are operating the ferroelectric material through half-switching (depolarized to polarized), and not full switching. However extreme values of CBL calculated for two-level readily agree with that for three level hence no such replacement is necessary.

For conventional bi-level FeRAM it is advantageous to have the maximum possible readout voltage. And conventionally the cell size is decided first depending on the technology of implementation and the capacity of the chip. Bitline capacitance is adjusted according to the C_{BL}/C_{FE} ratio that produces the maximum readout voltage. There is a considerable controversy in the literature about this ratio: in [13] this ratio is given as 1:1 or 1:2 while in [11] it is given as 1:10. This might be due to the particular technology and FeCap specifications chosen as well as the sensing scheme and the circuit techniques employed. However, in the case of a three-level cell we have to choose a proper value of C_{BL}/ C_{FE} so that there is a maximum signal difference between all the levels (not just V_H and V_L). Basically V_{SE} should be twice as big as before to keep the signal margins the same for 3 levels as they were for 2 levels. This allows sufficient margin between the consecutive[17] levels for better sensing. The margin has to be big enough (100-140mV) [21] so that the inevitable variations in device parameters (e.g. cell size, bitline capacitance, sense amplifier offset etc.) will not cause sensing errors. As well, as in case of conventional FeRAM, fatigue and relaxation considerations have to be taken into account during design to ensure reliable operation even after a large number of read and write cycles. Aging effect might impose severe limitations on the implementation of such three-level FeRAM. Considering the noise margins for these effects we also modeled the effects of parameter variations by including a distribution of possible bitline capacitances and cell areas [see Chapter 5].

---

[17] V_H-V_M and V_M-V_L

4.2 Read and Write Sequences

As mentioned in section 2.4.2, the read and write operations are governed by the ferroelectric capacitor characteristic hysteresis loop, which is in turn determined, by the material characteristic. A 'zero' corresponds[18] to a positive polarization of $+P_r$ (which in turn corresponds to a remnant charge of $+Q_r$) and a 'one' corresponds to a negative polarization of –Pr (which in turn corresponds to remnant charge of $-Q_r$). Now for a multilevel cell to have a middle level voltage, a third state needs to be defined in the characteristic hysteresis loop. To have the best noise margins between the two extreme voltages V_H and V_L, this third state should be exactly midway in between them. Hence the middle polarization state should be a zero polarization state. Achieving a zero (or even near-zero) polarization, however, is likely to be challenging because of the nonlinearity of the cell capacitors response to an applied voltage. This can be better illustrated with the help of Figure 4-4 from [11]. As shown in the figure, logic zero corresponds to point 'a' with a charge of Q_0 with respect to a negatively saturated state 'F', and point 'd' corresponds to a logic one with a charge of Q_1. Hence, in a read operation when the plate line is pulsed and the previous state is a one, there is a significant amount of charge switching from Q_1 to Q_0 (denoted by $2Q_r$), while if the previous state is a zero there is only a small amount of charge switching of Q_0. This directly implies that we will need a charge switch of Q_r to achieve the middle level signal voltage.

Assuming that the hysteresis loop is symmetrical about the X-axis, points 'e' and 'b' will have the same zero remnant polarization corresponding to point g. This implies that to write V_M we will need to write this voltage (V_M) on the BL and then pulse the plateline. However, we observed that simply writing V_M on the bitline does not ensure the desired operation of the cell because the next state depends on the previous state.

---

[18] Depending upon the convention used, $+Q_r$ might represent a '1' or '0' and $-Q_r$ complement of that.

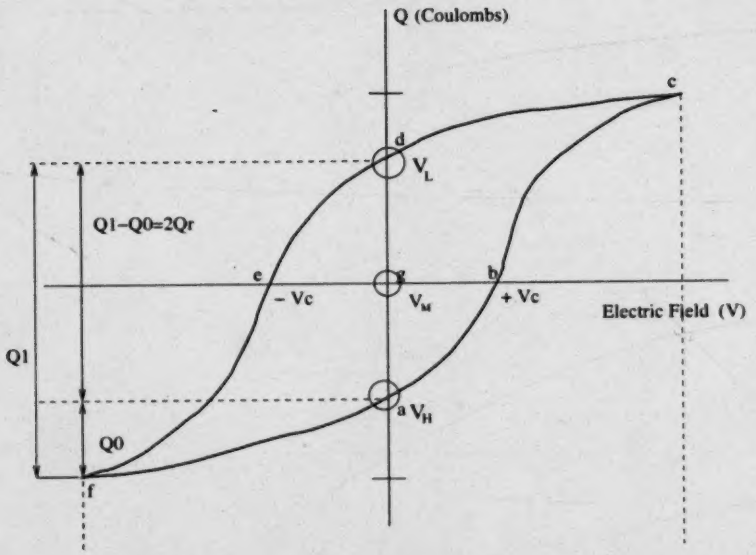
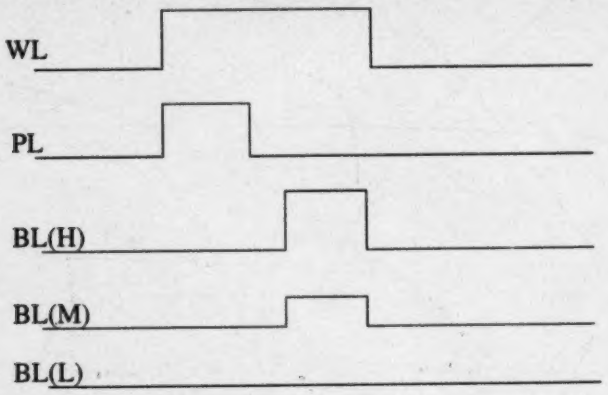


Figure 4-4 Ferroelectric Capacitor Characteristics [11]

We observed that, although the material reaches the state corresponding to V_M, the voltages read for a cell in this state gave different readout voltages depending on the previous state of the material. This was also true in case of state one and zero when the previous state is the state corresponding to V_M. Hence it was identified that special precautions are needed which will ensure correct read voltages for all the states independent of the previous state of the material. It is immaterial here if we achieve this with a read or a write since in FeRAM, the read (destructive) and write cycle times are equal.

The write sequence was modified as shown in Figure 4-5. With this sequence of operations we achieve first a known state (state "0" in the figure) and then write the required voltage to go to the desired written state. This ensures a constant readout voltage for each state that is independent of the previous state. Figure 4-6 shows various combinations of states achieved with the help of black dot, which traces through the hysteresis loop. The read operation of the multilevel cell remains the same as conventional two-level FeRAM. However the restore

operation (which is a write) changes. Figure 4-7 shows the hysteresis trajectories for multilevel read operation. An additional complexity of sublooping (discussed in Section 4.3) technique could be employed for restore operation (to write V_M).



(a) Method I

Figure 4-5 WL, BL and PL Waveforms for Write Operations.

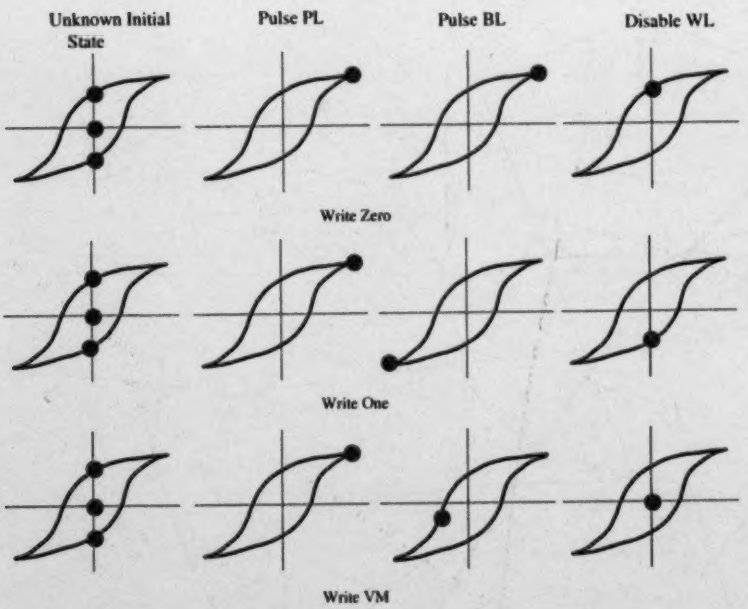


Figure 4-6 Hysteresis Trajectories During ML Write Operations

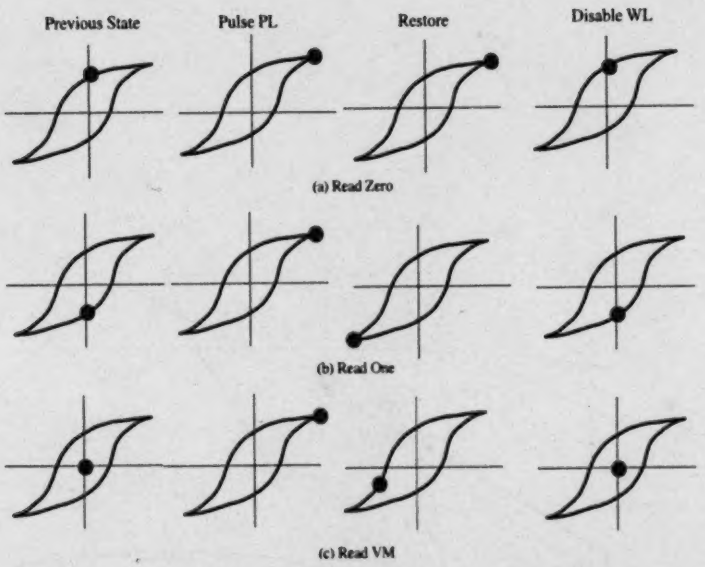


Figure 4-7 Hysteresis Trajectories During ML Read Operations

At the architectural level the read operation is carried out by accessing two cells (each storing one of the three levels) at a time. Hence we have 3 levels (from the first cell) x 3 levels (from the second cell) = 9 combinations. Only eight of those combinations are required to encode the state of three bits. Extra decoding logic is involved to translate the combinations of two cell signals to triplets of binary values. Hence we have 3 bits / 2 cells, i.e. 1.5 bits per cell. The biggest technical challenge is to determine an accurate way of writing the middle level voltage. This difficulty arises out of the fact that the polarization characteristic of the ferroelectric capacitor is a nonlinear function of the applied voltage and is steep near the point where the polarization goes to zero.

4.3 Design of the Middle Level Voltage Driver

Designing a middle level voltage driver is a crucial step in obtaining a working MLFeRAM. On-chip voltage generation is becoming increasingly important in

memory designs, especially with the recent trend towards lower-voltage operation [25]. Modern commercial memories incorporate on-chip voltage down-converters as well as voltage boosters based on charge pumps. Common techniques, like the capacitor charge sharing technique used in MLDRAM for writing the middle level voltage, cannot be used for programming MLFeRAM. This is due to the fact that a strange intermediate voltage is required from a low impedance driver, which has to drive a highly capacitive bitline and highly capacitive cells. Hence the design requirements for the middle level voltage driver are quite stringent. The driver should have high *power supply rejection ratio (PSRR)*, high current drive, low output impedance and high input impedance. As well it should be immune to noise, temperature changes and process variations. We first explored the possibility that these characteristics could be met by using an operational amplifier.

4.3.1 Method I

An operational amplifier (OPAMP) was designed to be used as a voltage follower for writing the middle level voltage onto the FeCap. Figure 4-8 shows the schematic of a two-stage CMOS OPAMP [35]. The first stage is a high gain differential amplifier input stage. The gates of M1 and M2 receive the differential input signal. M5 acts as an adjustable current source. M3 and M4 form a current mirror and act as a load. The second stage is a common source stage, which converts the second-stage input voltage signal to a current. Transistor M7 is loaded by a current-sink load [M6], which converts the current signal to a voltage at the output. C_C is a compensation capacitor used to avoid ringing at the output. Most OPAMPs used in feedback configurations might show a regenerative effect around the loop leading to instability or ringing or oscillation at the output. Various compensation techniques are employed to avoid this [35]. Miller capacitance compensation is very common and is designed with the help of a frequency response analysis [35]. Some of the important characteristics of the OPAMP are discussed here.

The common mode rejection ratio (CMRR) is defined as the ratio of the differential voltage gain to the common-mode voltage gain. The higher the value of the CMRR, the better the matching between the two input terminals and smaller is the output common-mode voltage. A smaller output common-mode voltage implies that the OPAMP has a better ability to reject common-mode voltages such as electrical noise, temperature changes etc. As derived in [35] we have:

$$CMRR = 2gm_1 gm_3 r_{o5} r_{o1} \tag{8}$$

where gm is the transconductance and r_o is the output resistance.

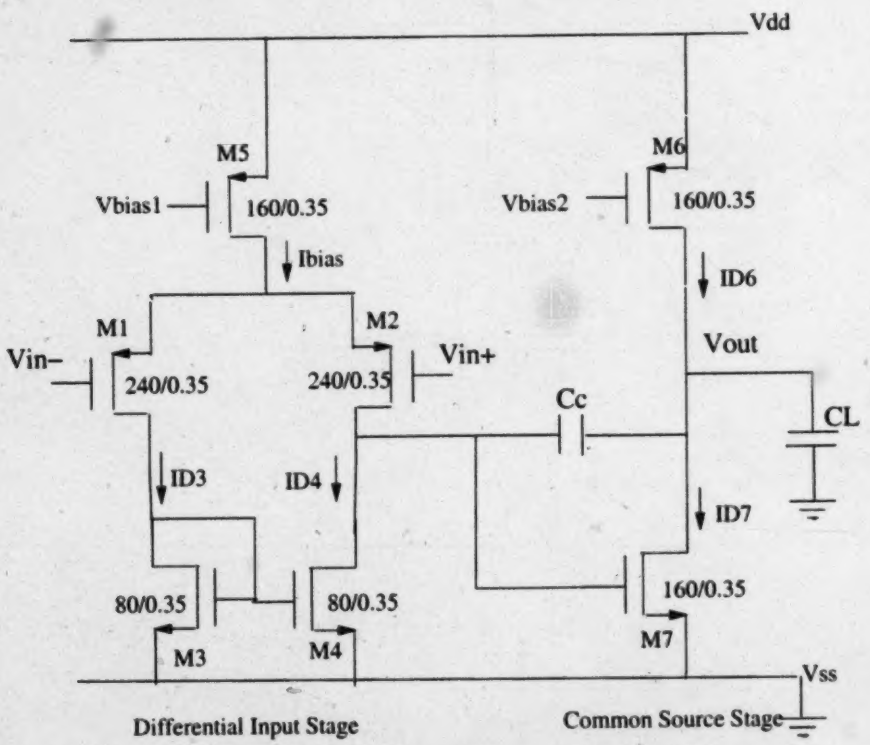


Figure 4-8 Two-stage Operational Amplifier

The slew rate is defined as the maximum rate of change of output voltage per unit of time and is usually expressed in volts per microseconds. The slew rate is an

important parameter for high frequency AC applications. For our design we are not as much interested in the slew rate as we are using this OPAMP as a voltage follower for a DC application. However the slew rate determines the rising time of the first overshoot and eventually the settling time of the output for a step input transient response. If we apply a large differential signal, the bias current $|I_{D5}|$ is intercepted by M2 and then charges C_C. Hence the slew rate will be given as in [42] by:

$$\text{Slew Rate} = S_R = \frac{|I_{D5}|}{C_C} = 2\frac{|I_{D1}|}{C_C} \tag{9}$$

$$A_{v1} = g_{m1}(r_{o2} \| r_{o4}) \tag{10}$$

$$r_{o4} = \frac{1}{I_{D4}\lambda n} \tag{11}$$

$$r_{o2} = \frac{1}{I_{D2}\lambda p} \tag{12}$$

$$g_{m1} = \sqrt{2\mu pCox\frac{W}{L}|I_{D1}|} \tag{13}$$

$$A_{v2} = -g_{m7}(r_{o6} \| r_{o7}) \tag{14}$$

The settling time of the operational amplifier is composed of nonlinear and linear regions. The time spent in the non-linear region depends on the slew rate of the design. The linear time is usually easier to obtain from simulation results.

The change in an OPAMP input offset voltage V_{io} caused by variations in the supply voltages is called the Power Supply Rejection Ratio (PSRR). The higher the PSRR the better is the OPAMP's immunity against power supply variations. PSRR is sometimes separately specified for the positive and negative supply voltages.

$$PSRR^{+} = \Delta V_{DD}/\Delta V_{IOS} = Av/A_{VDD} \tag{15}$$

Similarly $PSRR^{-} = Av/A_{VSS}$

Values of A_{VDD} and A_{VSS} are derived from small signal analysis and are as shown below:

$$A_{VDD} = r_{o7} * \frac{r_{o7}}{r_{o6} + r_{o7}} - gm7 \frac{r_{o6} * r_{o7}}{r_{o6} + r_{o7}} * \frac{1}{2 * g_{m3} r_{o3}} \quad (16)$$

$$A_{VSS} = \frac{r_{o6}}{r_{o6} + r_{o7}} \quad (17)$$

The differential amplifier described above was connected as a voltage follower as shown in Figure 4-8. V_{REF} is the voltage, which could be derived from a simple resistive ladder network. Vout stabilizes to V_{REF}. The sizes of C_C and C_{BL} are adjusted to get appropriate settling time and transient response of the output.

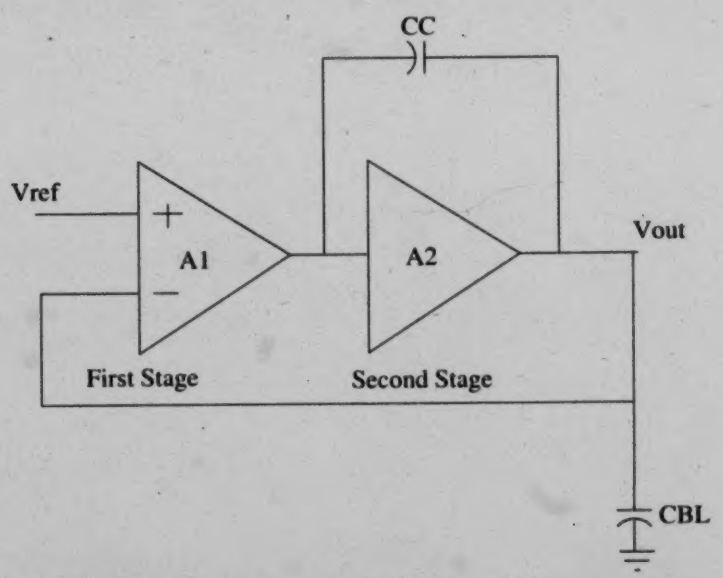


Figure 4-9 Block Diagram of OPAMP as Voltage Follower

One of the benefits of negative feedback is the stabilization of the voltage gain of an amplifier against changes in the component values (e.g., with temperature, frequency, etc.). The settling accuracy of the output voltage is governed by that of the reference voltage V_{REF}. V_{out} is almost equal to V_{REF}, however for a higher V_{REF} its accuracy against parametric variations may not be satisfactory [25].

Although a driver designed in this way is quite stable in generating V_M, depolarizing the material to middle "Zero State" is not always achieved. By depolarizing to the zero state we mean reaching point 'Y' in Figure 4-10. To illustrate this further, assume that we are in the process of writing V_M and we have reached point 'Q'. Now, due to cell size variations (or equivalent process mismatches), while we sweep back from $+V_c$ to 0V, the shrinking polarization might follow three[19] different paths Q-X, Q-Y or Q-Z on the nonlinear hysteresis loop and reach three different remnant polarizations at 'X', 'Y' and 'Z'. A similar variability in polarization will result while sweeping from $-V_c$ to 0V as shown by paths P-X, P-Y and P-Z. This may lead to different readout voltages from the cells and hence a wider distribution (discussed in section 5.3.2) of V_M. The ultimate result of the scatter in V_M is smaller noise margins between the actual distributions of V_L, V_M and V_H signals in the cells.

One potential way to improve these noise margins and achieve a narrower distribution of V_M is by traversing through a spiral of inner subloops, as shown in Figure 4-11, and more gradually reaching the depolarized state at point 'Y' in Figure 4-10. The advantage of this is that the slopes of the inner subloops tend to be less steep. Also, the Pr's of the inner subloops are very close to zero polarization. The transient response of a step input to the voltage follower is a damped oscillation and therefore could be used to generate subloops in the hysteresis loop. This is obtained by holding the plate line at the middle level voltage and applying the damped oscillating waveform onto the BL. Refer to Figure 4-12 for such waveform. This method achieves a better (narrowed) distribution for V_M in the presence of cell parameter variations[20].

---

[19] In fact there will be several such paths, leading to several different remnant polarizations.

[20] Discussed in section 5.3

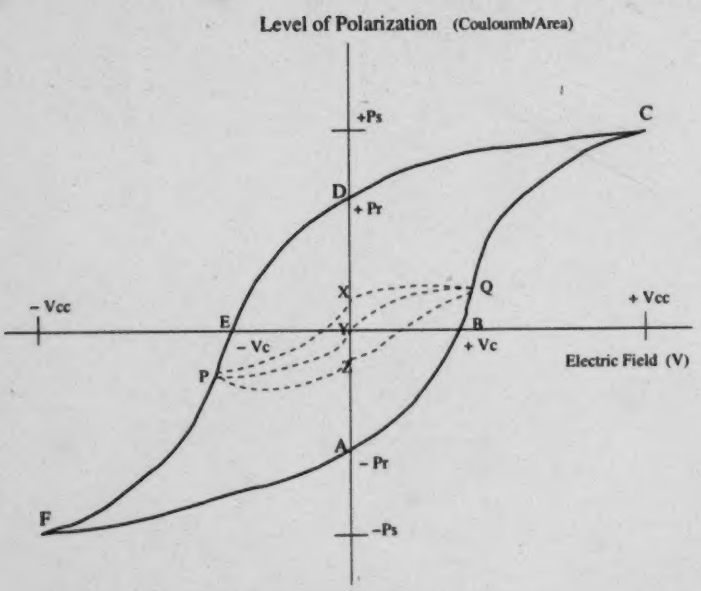


Figure 4-10 Depolarizing the Material to the Zero Polarization State

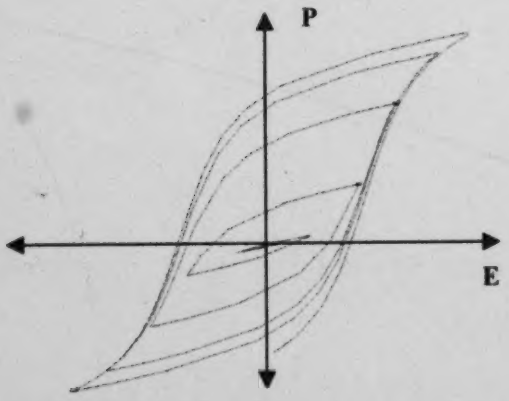


Figure 4-11 Reaching the Zero Polarization State Through Spiral Inner Subloops

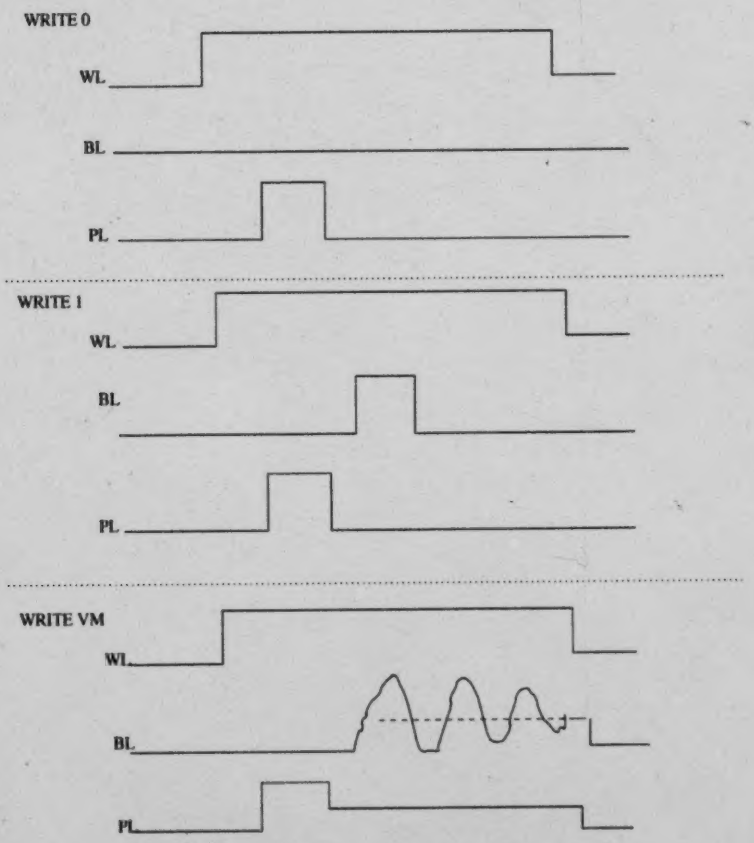


Figure 4-12 Write Control Waveforms

Design Specifications

The design of a suitable OPAMP bitline driver starts with the assumption that all of the transistors are appropriately biased to operate in the saturation region. An I_{Bias} of 100μA was chosen which makes both I_{D1} and I_{D2} equal to 50 μA. I_{Bias} is also drained through M3 and M4 giving $I_{D3} = I_{D4} = 50\mu A$. Also $I_{D7} = I_{D6} = 100$ μA. These values were found to be a good starting point for the design. After several design cycle iterations the sizes of the transistors were adjusted to give the best gain and best values for important characteristic like PSRR and CMRR, etc.

Table II SPICE BSIMV3 Model Parameters for the CMOS 0.35-μm TSMC Process

Spice Model Parameter	Symbol [Units]	Value
Oxide Capacitance	Cox [fF/μm^2]	~5
Carrier mobility of Electrons Holes	 μn [cm^2/V/s] μp [cm^2/V/s]	 ~400 ~100
Channel Length Modulation n-channel p-channel	 λn [V^{-1}] λp [V^{-1}]	 ~0.097 ~2.213
Threshold Voltage n-channel p-channel	 $\|V_{THN}\|$ [V] $\|V_{THP}\|$ [V]	 0.59 0.72
Gate Oxide Thickness	Tox [m]	7.5e-9

This led us to select the following device sizes[21]:

$$(W/L)_1 = (W/L)_2 = 240/0.35,\ (W/L)_3 = (W/L)_4 = 80/0.35,$$

$$(W/L)_5 = (W/L)_6 = (W/L)_7 = 160/0.35,$$

In general the larger the sizes (i.e., larger W/L ratio) of the MOSFETS that we used for the design in the differential pair, the higher will be the gain and the smaller will be the Vgs voltage. This increases the CMRR, increases the PSRR, lowers the input noise and hence better matching is achieved. The following calculations were made following the design procedure described in [35]:

First Stage Gain

$$r_{o4} = \frac{1}{I_{D4}\lambda n} = 206.83\text{K}\Omega,\ r_{o2} = \frac{1}{I_{D2}\lambda p} = 8.96\ \text{K}\Omega,\ r_{out1} = r_{o2} \| r_{o4} = 8.59\ \text{K}\Omega$$

[21] The OPAMP designed might not be optimum in terms of area, as the main goal considered here was to achieve better transient response and sensitivity parameters.

$g_{m1} = \sqrt{2\mu pCox(\frac{W}{L})_1 \mid I_{D1} \mid}$ = 1.9 mA/V, $A_{V1} = g_{m1} * r_{out1}$ = 15.9164

Second Stage Gain

$r_{o7} = \frac{1}{I_{D7}\lambda n}$ = 103.41 KΩ, $r_{o6} = \frac{1}{I_{D6}\lambda p}$ = 4.48 KΩ, $g_{m7} = \sqrt{2\mu nCox(\frac{W}{L})_7 \mid I_{D7} \mid}$ =

4.3mA/V, $A_{v2} = -g_{m7}(r_{o6} \parallel r_{o7})$ = 18.3787, Av = $Av_1 * Av_2$ = 292.5242

CMRR

$g_{m3} = \sqrt{2\mu nCox(\frac{W}{L})_3 \mid I_{D3} \mid}$ = 2.1mA/V, $r_{o5} = \frac{1}{I_{D5}\lambda p}$ = 4.48 KΩ

$CMRR = 2gm_1 gm_3 r_{o5} r_{o1}$ =318.4438 ≅ 50.0607dB

PSRR

$PSRR^+ = Av/A_{VDD}$ = 2.6348e+018 ≅ 368.4150dB

$PSRR^- = Av/A_{VSS}$ = 7.0384e+003 ≅ 76.9495dB

Slew Rate

$S_R = \frac{\mid I_{D5} \mid}{C_C} = 2\frac{\mid I_{D1} \mid}{C_C}$ = 100 V/μs

Transient Response

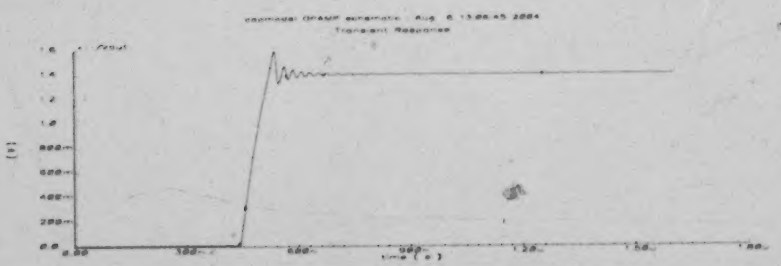


Figure 4-13 Transient Response for a Step Input

The settling time of the OPAMP driver can be found by simulating the step response. The transient response depends on the bitline capacitance C_{BL} and the compensating capacitance C_C. The damping factor depends on the ratio of these two capacitors as well as on the gain of the operational amplifier. Parametric analysis was run for different values of C_{BL}, C_C and A. The best transient response (with steeper rising time, minimum settling time and wider output swings) applied across the ferroelectric capacitor model was obtained using the following parameter for writing V_M:

A ≅ 300, C_{BL} =1pF-5pF, C_C = 500 – 800fF.
Rise time = tr ≅80ns, Settling time = ts ≅120ns
Peaks (above V_M) P1≅ 600mV, P2 ≅ 400mV, P3 ≅250mV, P4 ≅120mV, P5 ≅ 40mV
Troughs (below V_M) T1≅400mV, T2 ≅300mV, T3 ≅200mV, T4 ≅ 80mV, T5 ≅ 10mV

The difficulty with this method is that the output voltage swings obtained above and below the middle level voltage (i.e. fixed PL voltage) are rather small. As well there is a voltage drop (voltage drop across Ron) across the cell access transistor and the transistor switch used to isolate the OPAMP from the BL. This last transistor is only switched on while writing V_M onto the BL. Various techniques were tried to eliminate this transistor switch. However, eliminating this transistor creates a loading effect (when writing V_L and V_H) due to the output impedance of the OPAMP. We observed that this series resistance leads to serious shrinking of the hysteresis loop, i.e. the applied voltage swing is so reduced that it is unable to force the cell dielectric into saturation.

4.3.2 Method II

Another method of generating the oscillating PL voltage waveform is to use a digital-to-analog converter. Recall that the objective of generating such a waveform is to get inner subloops which will iteratively polarize the cell dielectric

material to a close to fully depolarized state, irrespective of the cell size and process variations. The goal was to achieve at least three such subloops and then return to point 'Y' in Figure 4-10. This could be done in two ways, as shown in the waveforms in Figure 4-14. In part (a) both the plateline and the bitline are pulsed so a differential voltage of V_{BL}-V_{PL} appears across the capacitor. In part (b) only the bitline is pulsed and the plateline voltage is held constant at half-Vdd. However, there are certain issues to be considered. Both of the waveforms will require the so-called BL||PL architecture with the BLs running in parallel with the PLs. The problem with waveform (a) is that the PL has to be pulsed to several intermediate voltages between V_{DD} and V_{SS}. Since the plateline is highly capacitive, pulsing the plateline in such a manner would cause a long time delay. Hence waveform (b) is preferred over (a). An effective way of generating such a waveform might be to use an on-chip or off-chip DAC. However here we have proposed an on-chip DAC, as on-chip generators tend to have better noise performance [32]. There are several standard techniques for implementing such a DAC. The resistor ladder method is simple but poor matching of the CMOS resistors renders it unsuitable for on-chip use [35]. Another approach is to use charge sharing among sized capacitors. The method is again relatively simple however once again low tolerances of CMOS capacitance rules it out. Moreover 2^n capacitances need large area [35].

A more advanced DAC is a current steering DAC in which an array of current sources is used to generate an analog sum current, which can be converted into a voltage via a current-to-voltage converter. The analog output voltage is a function of the number of current cells, which are switched according to the digital input code [34]. The advantage of this architecture is excellent DNL/INL[22] performance with thermometer-coded input. As well the design could be segmented to save area. Figure 4-15 shows the block level schematic of a current-steering DAC. The advantage of the current-steering DAC is that it can switch the

---

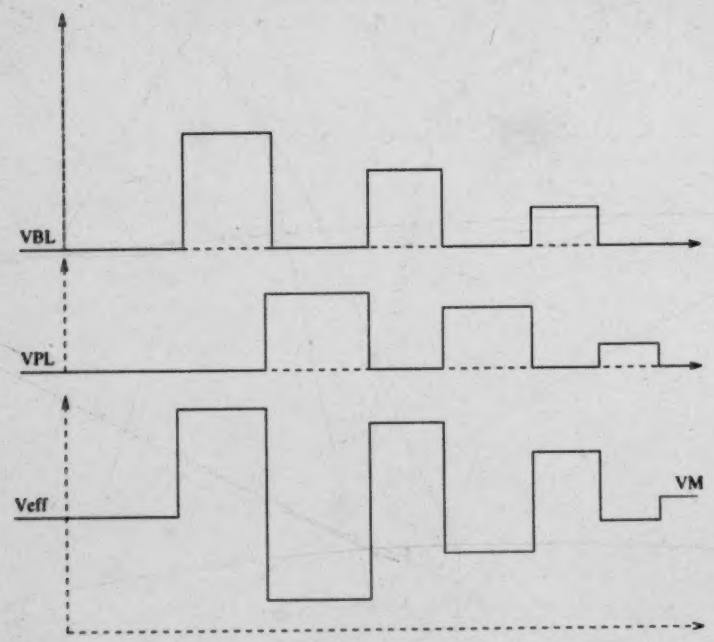
[22] Differential nonlinearity (DNL) is a measure of deviation of actual step size from ideal step size. Incremental nonlinearity (INL) is defined as the measure of deviation of the midpoint of the actual step form the midpoint of the ideal step [31].

current sources very quickly leading to fast D-to-A conversion [34]. This DAC is very often used in high-speed designs. Unfortunately current-steering DACs are rather sensitive to device mismatch [33]. Mismatch errors arise due to inevitable inaccuracies during chip processing. To improve the local matching of the DAC current weights, unit current sources are used in parallel to form the current weights. Hence for the *i*th LSB , 2^{i-1}LSB current sources could be used in parallel. Special layout techniques are needed for improved global matching. Due to high current distribution between the load impedance and the output impedance of the DAC, the linearity of the DAC will be affected. Hence the output impedance of the DAC has to be high. Nonlinearity arises in current-steering DACs from mismatches in the current-steering array and nonlinear characteristics of the trans-impedance amplifier or voltage dependence of the resistor for the current-to-voltage conversion [34]. A cascode transistor is typically used to increase the output impedance, which is roughly 100 times more than without it. Glitches arise due to mismatches between the switching times of the different bits: for short periods of time incorrect codes will be generated yielding large current spikes [34]. To reduce such error glitches it is common to use so-called segmented structures where a number of most significant bits are thermometer coded.

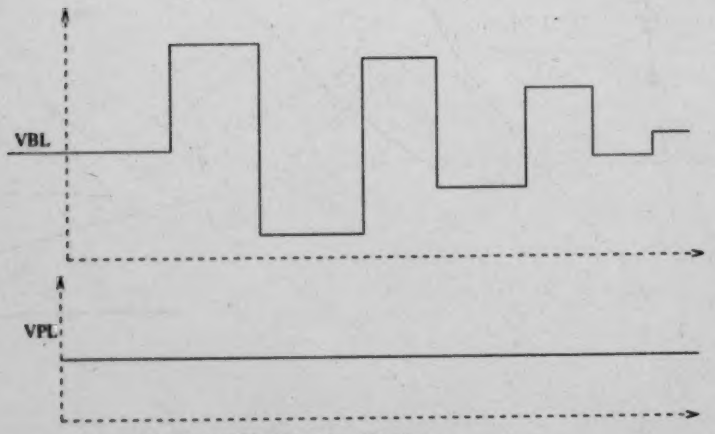
$$V_{out} = -R_F(I_0+I_1+I_2+I_3.....+I_{N-1}) \tag{18}$$

$$I_1 = 2\,I_0,\ I_2 = 2\,I_1,\ I_3 = 2\,I_2.....I_{N-1} = 2\,I_{N-2} \tag{19}$$

As well we can no longer rely on the output impedance of the current sources. Hence an OPAMP is employed which also acts as a current-to-voltage converter. Other advantages of an OPAMP include a higher common mode rejection ratio, a higher power supply rejection ratio and higher output impedance. Designing an OPAMP is an iterative process. To speed up the design time for the DAC, our OPAMP was modeled in Verilog-AMS with the required characteristics. Appendix A shows the code used for this design.

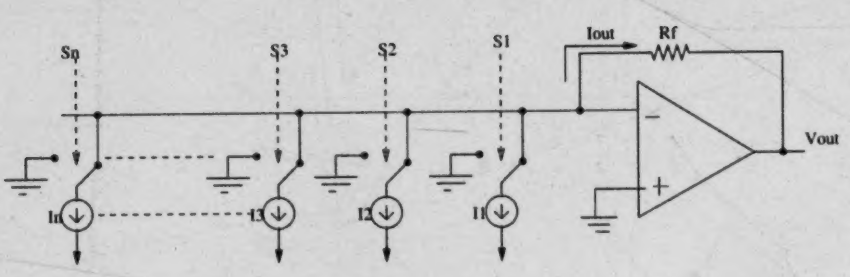


(a) Writing VM differential pulsing for BL and PL

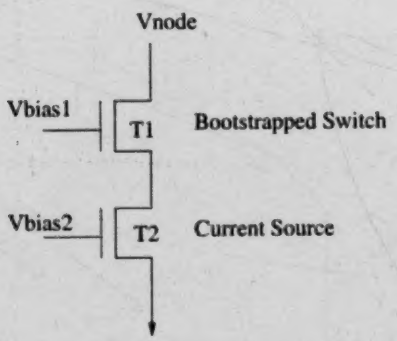


(b) Writing VM with constant PL voltage

Figure 4-14 Writing V_M with (a) Differentially Pulsing BL and PL (b) Constant PL Voltage



(a)



(b)

Figure 4-15 (a) Current Steering DAC (b) Single Transistor Current Source

Table III Operational Amplifier Characteristics

Characteristic	**Symbol**	**Value**
Input Resistance	rin [Ohms]	1e6
Gain	Gain	835e3
Slew Rate	slew_rate [A/F]	0.5e6
Output Limit Parameter	vsoft [mV]	200
Input Bias Current	Ibias [A]	0
Input Offset Voltage	Vin_offset [V]	0
Output Resistance	Rout [Ohms]	80
Unity Gain Frequency	freq_unitygain	1.0e6

A simple cascode was used as a current source as shown in Figure 4-15 (b). Transistor T1 acts like a switch and is bootstrapped. Transistor T2 is used as a

simple current source. It has to be appropriately biased to work in the saturation region with a drain current of:

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{TH})^2 \qquad (20)$$

Usually the OPAMP output range (or swing) is limited by the output stage of the OPAMP. In our design the output swing was limited by 200mV from the rails. Designing a DAC that has a rail-to-rail output voltage swing is a challenging task. The full scale range (FSR) is the difference between the maximum and minimum analog values and is equal to 3.18 - 0.2 = 2.98 V in our design. The design decision of I_0 to be around 1.972 μA was made on the basis of this. The currents I_1 I_2 ....I_{N-1} were obtained as in (19). R_F of 80K was thus obtained. Figure 4-16 shows the input-output characteristics of the implemented 5-bit DAC. As can be seen, the output characteristic is quite linear until 3.0124V, which is equivalent to a digital input code of 10101. Advanced issues, like calibrating the current sources (by resizing the transistors), analysis of non-idealities like switching errors, leakages, biasing, linearization techniques, were ignored as they were out of the scope of this project.

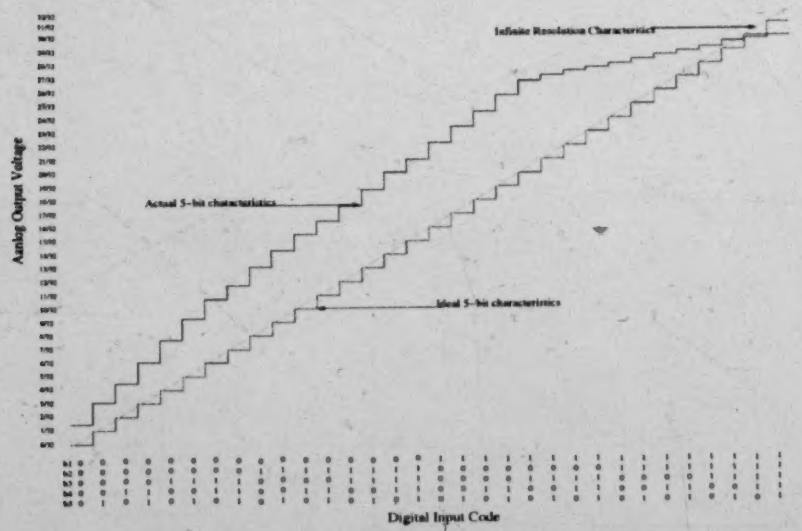


Figure 4-16 Input-output Characteristics of the 5-bit Current Steering DAC

4.4 Reference Voltage Generation

Several different reference voltage generation techniques have been proposed for 1T1C cell structures [Section 2.4.1]. The multilevel memory cell concept imposes more stringent requirements on accurate reference voltage generation than bi-level memory cell. Also, the reference voltage generated should be adaptive to the actual distribution of cell polarizations. Adaptive references could compensate for changes in cell characteristics caused by endurance, fatigue and imprint in the ferroelectric material. These characteristic variations lead to shifts or variations[23] in the readout signals of the cells. Ideally ferroelectric reference cells (capacitors) should be used to automatically track the process variations and material characteristics, which might vary throughout the lifetime of the cell. However, as discussed in section 2.4.1, these reference capacitors would then tend to fatigue more than regular data cells and would thus generate inaccurate reference voltages leading to early memory failure. In the present work an auto-calibrated reference voltage generation scheme is proposed to track the variations in the signal voltage levels of the cells. This has been implemented for three-level operation of the cell.

4.4.1 Memory Core Floorplan

Figure 4-17 shows a block-level view of the memory array floorplan. As in any memory, the array contains a two-dimensional grid of intersecting BLs (rows) and WLs (columns). PLs are also present running parallel to either BLs or WLs. SAs are usually placed at one end of BL and BLN pairs. However in this design, since we need to perform two comparisons (this will be evident soon) we need SAs at both the ends to obtain fast parallel sensing. Added to this, a precharge block is introduced at both the ends of the memory core. The memory core (array) is internally partitioned into several sections, with each section consisting of 256 bitlines. It is explained later why the number 256 was chosen for this partitioning.

---

[23] Refer to section 2.5.3 in Chapter 2 to understand how these variations occur due to several effects

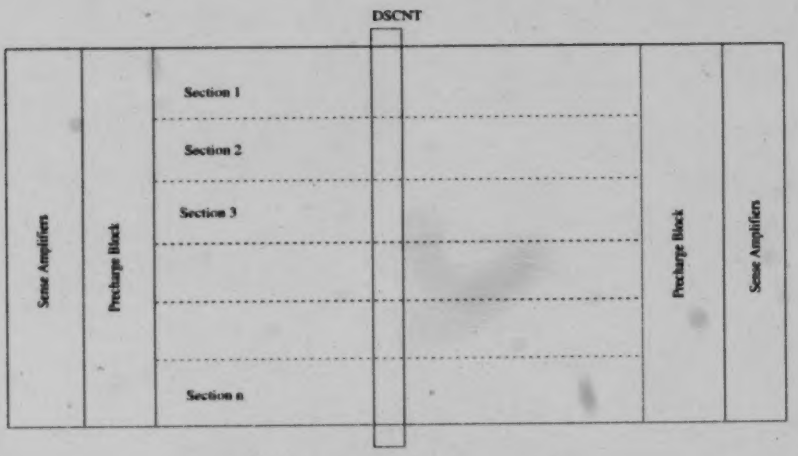


Figure 4-17 Memory Core Floorplan

4.4.2 Array Architecture

Figure 4-18 shows a detailed view of one section in the memory array. Sense amplifiers and precharge blocks are connected at the ends of the memory columns. DCNT controls the n-channel pass transistors, which divide the BL (as well as the BLN) into right and left side sub-bitlines. CHSH_BL and CHSH_BLN control the charge sharing transistors from the common bus, which independently charge shares all the BLs and all the BLNs. Although not shown in the figure, CHSH_BL and CHSH_BLN are present in both sections. These blocks introduce an area overhead. Alternatively, precharge blocks could only be used at one end of the BL(BLN) to eliminate the need for separate CHSH_BL and CHSH_BLN in both the sections of the BLs. However the two references needed for the SAs would then need to be generated sequentially, thus imposing an access time penalty.

Figure 4-19 shows the internal structure of the precharge block. This block precharges the BL or BLN depending on signals PRE1 or PRE2, respectively. This precharge voltage is either V_{DD} , V_{SS} or half-Vdd depending on the bit stored in the flip-flop. Two such precharge blocks are used at the either end of the BLs by means of charge sharing over multiple BLs (or BLNs). They are used to

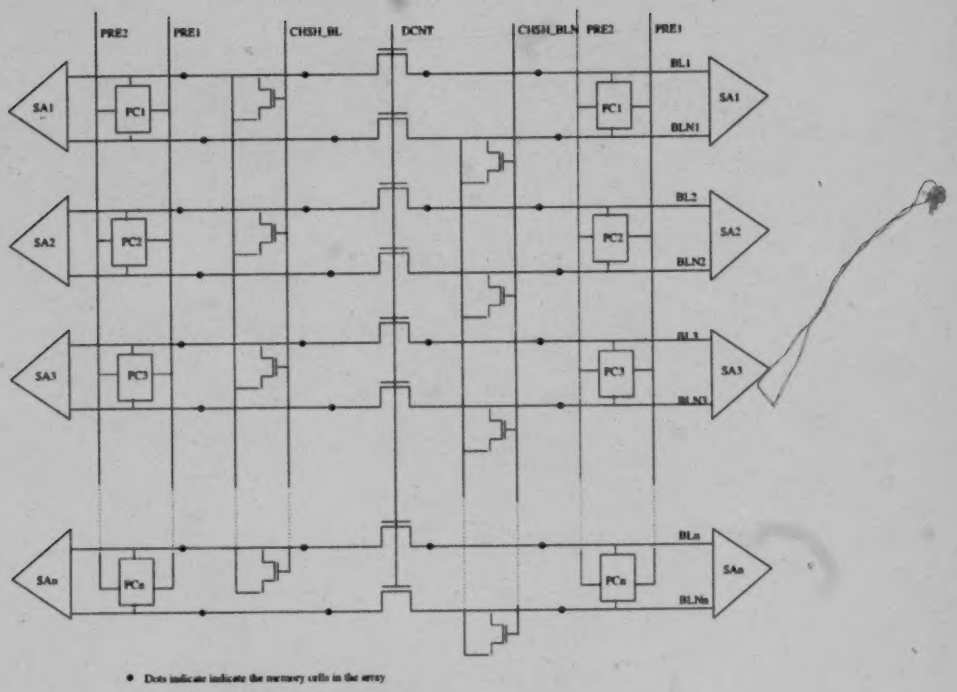


Figure 4-18 Architecture of an Array Section

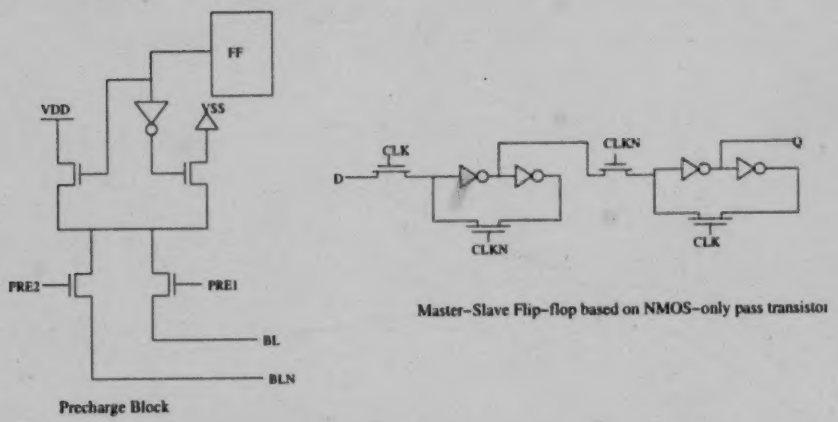


Figure 4-19 Internal Structure of the Precharge Block

generate V_{REF1} and V_{REF2} at the respective ends of the BLs. In the case of the three-level FeRAM, since V_{REF1} has to be in between V_{DD} and half-V_{DD} and VREF2 to be in between half V_{DD} and V_{SS}, the precharge voltages are half-V_{DD} and V_{DD} for the left block and half-V_{DD} and V_{SS} for the right block. Since 256 such bitlines are charge-shared together, a range[24] of decreasing voltages is available between half-V_{DD} and V_{SS} (between V_{DD} and half-V_{DD} for the right block). Table IV shows the range of values of V_{REF} resulting from charge sharing. To achieve this charge sharing, all the bitlines are connected to a common bus, which is precharged to half-V_{DD}. The number of BLs is flexible and could be changed to suit the need of design. Here we obtain a precision of 1.65/256 ~0.0064V. The flip-flop employed in the precharge circuit is a master-slave negative edge-triggered flip-flop, constructed out of two NMOS pass transistor latches. Another way is to use a static cross-coupled CMOS latch. The read operation is started by accessing a cell on the BL (or BLN). The readout voltage is available on the BL (BLN). At this point DCNT goes low and the BL (BLN) is partitioned into two sections containing two copies of the signal voltages. Following this PRE2(PRE1) goes high and all of the BLNs(BLs) are precharged to a predetermined voltage. When PRE2(PRE1) goes low CHSH_BLN(CHSH_BL) goes high thus charge sharing all of the 256 BLNs(BLs). This generates the required reference voltage, which is fed to one end of the SAs. However the challenge here is to pitch match the layout of precharge block and the sense amplifier within the specified minimum pitch of the BL.

[24] 256 different values in this design

Table IV Spectrum of the Values of V_{REF1} and V_{REF2}

No. of BLs precharged to half-Vdd	**No. of BLs precharged to Vdd/Vss**	V_{REF1}= (n*halfvdd+(256-n)*vss)/256	V_{REF2}= (n*halfvdd+(256-n)*vdd)/256
256	0	1.65	1.6500
255	1	1.6436	1.6564
254	2	1.6371	1.6629
253	3	1.6307	1.6693
252	4	1.6242	1.6758
251	5	1.6178	1.6822
250	6	1.6113	1.6887
249	7	1.6049	1.6951
248	8	1.5984	1.7016
|	|	|	|
|	|	|	|
8	248	0.0580	3.2420
7	249	0.0516	3.2484
6	250	0.0451	3.2549
5	251	0.0387	3.2613
4	252	0.0322	3.2678
3	253	0.0258	3.2742
2	254	0.0193	3.2807
1	255	0.0129	3.2871
0	256	0.0064	3.2936

* Note that capacitance of the charge sharing line (~250fF) was not taken into consideration for these calculations; however, it could be easily taken into account for precise calculations.

4.4.3 Error Checking/Auto Calibration Mode

Since MLFeRAM has smaller noise margins than its two-level counterpart, it is more sensitive to any reference voltage variations and readout voltage variations. Hence as the characteristics of the cell dielectric change through the life of the data cells, the reference voltages should ideally automatically track the signal changes. Before any fixed values of reference voltages can be used for sensing, it may be desirable to have the chip re-calibrated. To achieve this the reference voltage could be swept through a range of values available from the charge sharing circuit. As explained before, charge sharing is done to obtain the required voltage by connecting *n* number of bitlines. The value stored in the flip-flop of the precharge circuit is changed to obtain various combinations of BLs being precharged to different voltages (half-V_{DD}, V_{SS} and V_{DD}). These flip-flops are connected in a chain to form an *n*-bit shift register. This eliminates the need of connecting the Din of each flip-flop to a separate external input.

The error checking process starts with writing a specific value to a cell and then reading it back for the given value of V_{REF1} and V_{REF2} at that time. This is repeated for all the cells in the memory core. The read value is compared with the written data to detect any errors. The same process is repeated for the next value of V_{REF}. A histogram is thus collected for the errors with respect to V_{REF}. The histogram takes the form of a bathtub curve. As shown in the Figure 4-20 the percentage of errors is maximum at the extreme values of V_{REF} while it reduces to zero for the middle range values. The two extreme points at which the errors start to build up are noted and the average of these two is taken as the reliable value of reference voltage. Then the chip is calibrated for this reference voltage. The binary control values are written in the flip-flops (precharge voltage values of the BLs). These preserved voltages in the flip-flops of the precharge blocks when charge shared give the required V_{REF1} and V_{REF2}. The calibrating process can be implemented as a state machine, which is part of peripheral logic or it could be implemented in an externally sequenced testing algorithm, which is external to the

chip and can be programmed as required. Another possible alternative is to store a fixed calibration setting in conventional FeRAM cells.

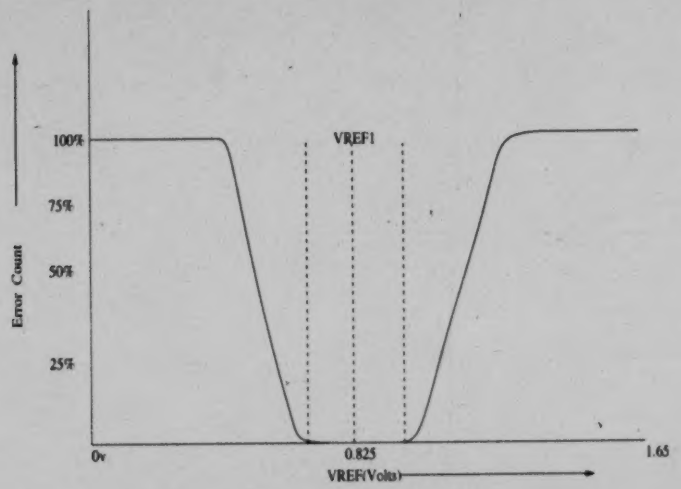


Figure 4-20 Bath-tub Curve Obtained by Sweeping the Reference Voltages

The proposed technique provides programmability in the reference voltage with relatively little area overhead within the array and without extra sensing time. Sensing time is not affected because auto-calibration of the reference voltage is done well before the actual read and write operations of the chip. Also charge sharing is accomplished well in advance when the read access is started. This technique is anticipated to result in better yield and robustness against minor cell parameter variations. Also the generated reference is stable (negligible fluctuations) against any power supply and temperature variations unlike conventional techniques.

Chapter 5

MLFeRAM Operation in the Presence of Parameter Variations

As discussed in [1] and seen from the various cell design graphs in the previous chapter, the area of the capacitor determines both the stored and switched charge and hence the signal strength on the bitline. Another important factor, which might determine the bitline signal strength before sensing, is the bitline capacitance C_{BL}. Recall that for a parallel plate capacitor $C=\epsilon_0 * \epsilon_r * A/d$. Here ϵ_r is the fixed relative permittivity for the given ferroelectric material and ϵ_0 is the permittivity of free space. Also, for a given ferroelectric process, the thickness d of the dielectric is fixed; hence A is the only tuning parameter here.

Typically in MLDRAMs we have noise margins of at least 50-100 mV between data signal levels. However for FeRAMs, due to reliability issues, we should target for higher noise margins in the range of 100-150 mV for better sensing. It is vital that the signals [3] have large enough noise margins in between them so that it's easier to differentiate the signals as V_M (Middle), V_H (High), V_L

(Low) and sensing errors are avoided. Also, a prior knowledge of the readout voltage drifts[25] gives some idea of the reference voltage generation. A conventional test [37] to achieve this in a real test chip is called a functional test, which is based on simply writing ones and zeros and then reading them back and checking for sensing errors. A charge distribution graph of the cells is plotted for given set of conditions. Reference [37] gives a brief idea about the charge distribution and also proposes an evaluation scheme for that. Figure 5-1 below shows the charge distribution of cells in [37]. As we can see the distribution is different for data 0 (corresponding to capacitor charge of Q_r) and data 1 (corresponding to capacitor charge of $-Q_r$) and that there is a clear gap of around 145 fC, which is considered good enough [37] for two-level operation. The scattered distribution of the cell charges is as a result of inevitable manufacturing process variations in the cell characteristics, plateline strengths, circuit parasitics, etc. An appropriate reference voltage should be placed in the middle of the gap when performing sensing (read) operations to minimize the probability of sensing errors. A common approach is to position the reference voltage exactly in between the two distributions. However, as we can see the data 0 distribution is narrower than that of data 1, hence the reference is sometimes placed closer to the tail of data 0 than that of data 1. One recent paper [39] presents a technique to measure the charge distribution for a 4-Mbit FeRAM. Reference [3] proposes a novel high-speed cell charge distribution measurement technique and on-chip compression of this distribution data. There are other advanced issues, like tracking variations in the cell characteristics, because the ferroelectric capacitors undergo various time-dependent phenomena, like fatigue, retention and relaxation that require some sort of programmability in the reference voltage. These techniques are summarized in [1].

[25] Caused by inevitable statistical parametric variations.

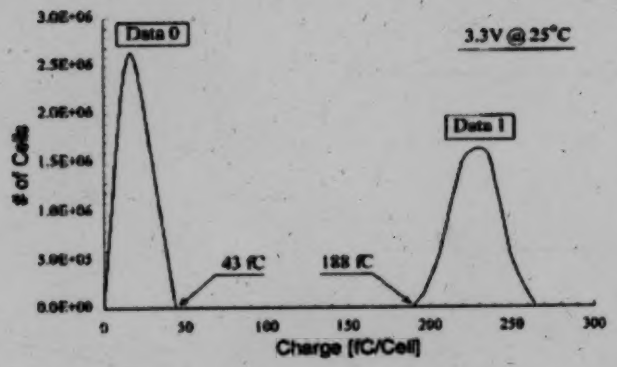


Figure 5-1 Cell Charge Distributions for Data 0 and Data 1 [39]

In the case of the proposed multilevel FeRAM, this sort of charge distribution couldn't be collected experimentally because a test chip wasn't fabricated. However, to better understand the distribution for three levels, various parameters were varied in simulation and the resulting distributions were plotted. Since cell area is probably the most important factor affecting the signal produced by a cell, the charge distribution curve was collected for up to +/-30% variation in area. The following approach was taken. Initially the cell area was varied by +/-30% about a mean area of $1\mu m^2$ in steps of 1%. The resulting signals were then weighted by an appropriately scaled normal distribution.

5.1 Schematic Simulation Setup

The design was implemented in Virtuoso Schematic Editor in the Cadence design environment with parameters derived from the CMOS 0.35-μm TSMC process available through CMC. The AHDL code for the ferroelectric capacitor model in [3] was edited to specify the new default parameters. A new symbol was created in Virtuoso with basic input output. Then a Cellview (AHDL) was created from this Cellview (Symbol). Then when we opened the library manager we got a file ahdl.def instead of a schematic. This file was edited by pasting the AHDL code into it. Later the symbol was edited to give all the respective pins in the model. This symbol now can be used for simulation and verification in Cadence. The

Appendix B shows the schematic in the Virtuoso editor. The simplest schematic is a FeCap with one access transistor connected to the bitline. The parasitic BL capacitance was modeled as a lumped capacitance connected to this BL. The Appendix B also shows the circuit that was used for observing the hysteresis loop in the Cadence Design environment.

The FeCap was connected across the voltage source of value 3.3 V, while the coercive voltage is +-1.5 V. The applied voltage is a triangular waveform, which varies from -3.3 V to 3.3 V. Because of the shunting effect, we cannot directly observe the polarization across the FeCap. Instead a linear capacitor was used to integrate the charge transferred across the FeCap. A voltage-controlled current source was connected across this capacitor. This current source is controlled by the voltage V_{fe} at the minus node of the FeCap model. Thus the variations in V_{fe} will develop variations in charge across the dielectric capacitor. Now since Q=CV for a capacitor, we can directly measure/plot V if C= 1F. However, the most convenient value of dielectric capacitor found for better observation of the hysteresis loop was 0.5 to 1pF. Hence the hysteresis loop obtained is the characteristics obtained with this V on Y-axis and the coercive voltage V_{cr} on the X-axis.

5.2 Review of the Normal Distribution

The normal distributions are a very important class of statistical probability density distributions [43]. All normal distributions are symmetric and have bell-shaped density curves with a single peak. Any normal distribution is specified with two parameters: the mean μ, where the peak of the density occurs, and the standard deviation, which indicates the spread or girth of the probability density curve. An important attribute of the standard deviation as a measure of spread is that if the mean and standard deviation of a normal distribution are known, it is possible to compute the percentile rank associated with any given score (or value). The standard deviation of the value is a better measure of probability than the range because it takes all the values into account. The more a score varies

from the average, the lesser the probability of its occurrence. The standard deviation σ, is the square root of the (biased) variance which in turn is defined as:

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} \tag{21}$$

where N is the total number of values considered.

The value of the *probability density function* (PDF) at any value x is given by

$$\text{PDF}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(x-\mu)^2/(\sigma)^2} \tag{22}$$

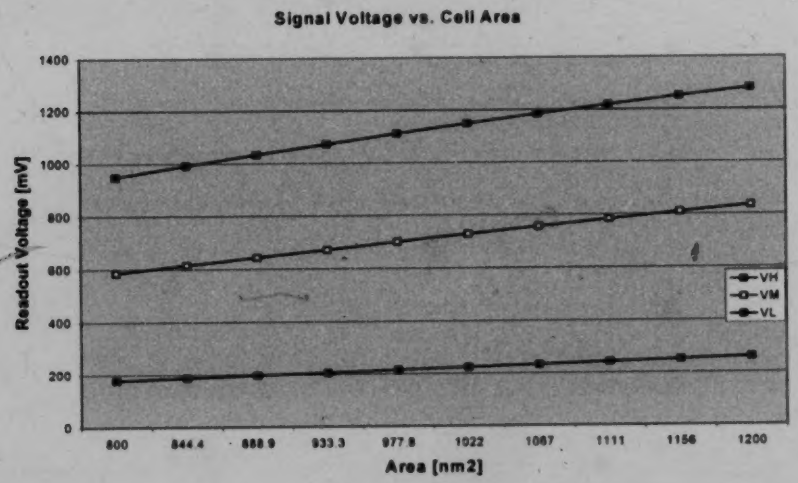
Figure 5-2 (a) Probability Distribution Function; (b) Cumulative Distribution Function

5.3 Parameter Variations

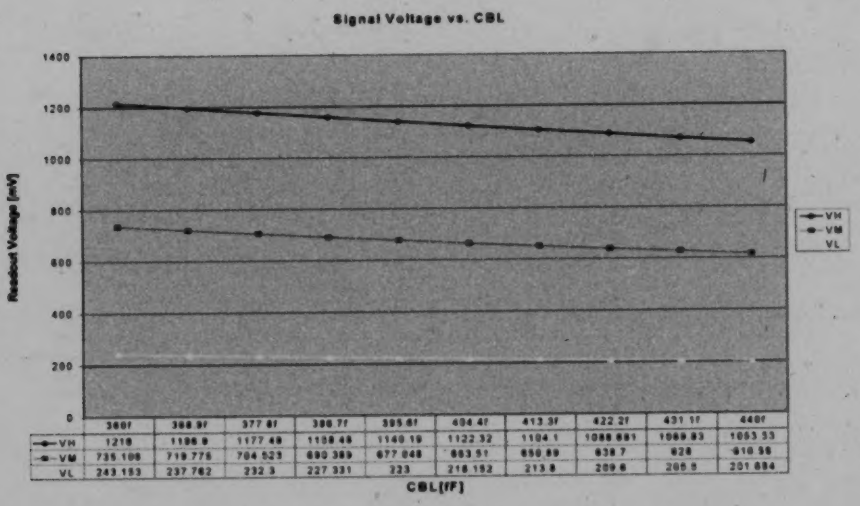
5.3.1 Process Variations

There are many different parameters that can affect the readout signal produced from the ferroelectric cell. The parameters related to the cell capacitance of the AHDL model are as follows: A = 1.0 um^2, d = 170 nm, ϵ_r = 350, Ps_ = 30μC/cm^2, Pr_ = 25μC/cm^2, V_{cp} = +1.5 V, V_{cn} = -1.5 V and R_{leak} = 5kΩ. The challenge is to model these parameter variations in a simulation environment. Since the design was implemented in 0.35-μm technology, the area of the capacitor plates was chosen to be 1.0 μm^2 as an appropriate value. As given in [37] for a PZT capacitor with A= 1.0 μm^2 and a Pr of 25 μC/cm^2, the remnant charge will lie

somewhere in the range of 200 – 300fC. Figure 5-3 and Figure 5-4 show the effect for variations in some of the parameters. The Y-axis shows the readout voltages while the X-axis show the parameter being varied.

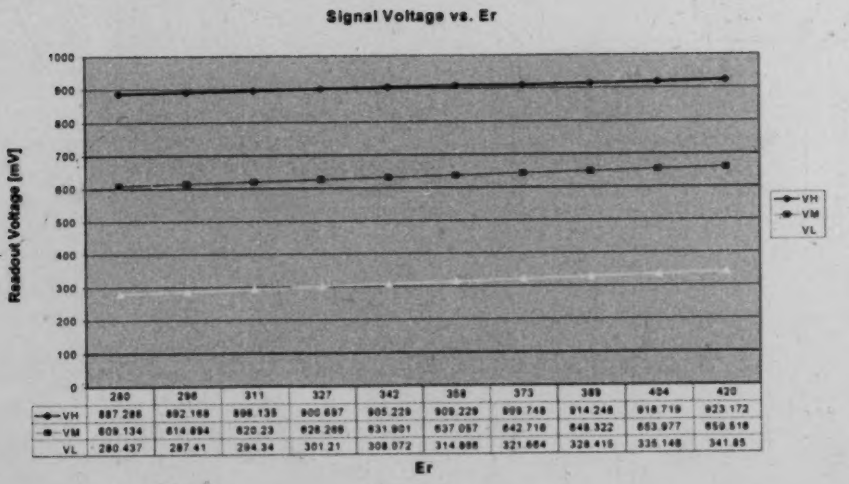


(a)



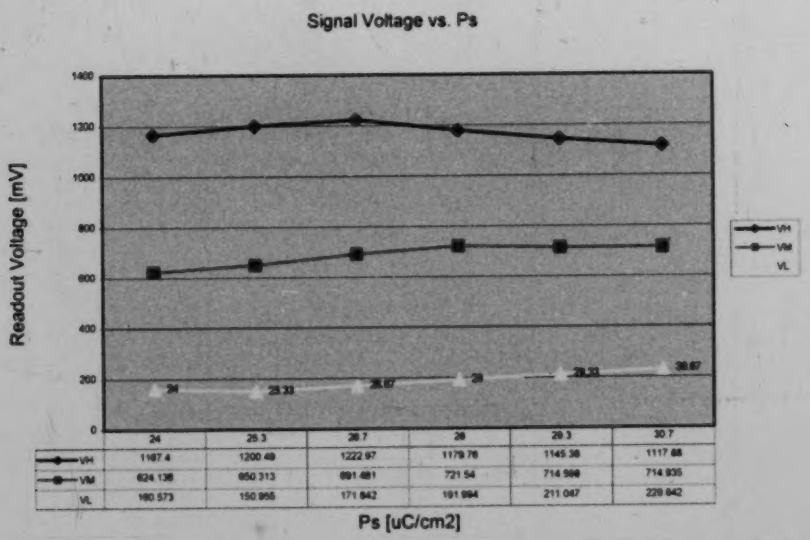
(b)

Figure 5-3 Readout Voltages with Variations in (a) Area and (b) C_{BL}



	280	296	311	327	342	358	373	389	404	420
VH	887.286	892.169	896.135	900.697	905.229	909.229	909.748	914.248	918.719	923.172
VM	609.134	614.894	620.23	626.269	631.901	637.057	642.716	648.322	653.977	659.518
VL	280.437	287.41	294.34	301.21	308.072	314.888	321.664	328.415	335.148	341.85

(a)



	24	25.3	26.7	28	29.3	30.7
VH	1187.4	1200.49	1222.97	1179.76	1145.38	1117.68
VM	624.138	650.313	691.481	721.54	714.998	714.935
VL	160.573	150.955	171.842	191.994	211.047	229.842

(b)

Figure 5-4 Readout Voltages with Variations in (a)ε_r and (b) Ps

5.3.2 Simulation Technique for Noise Margin Analysis

According to the *central limit theorem* the data, which are influenced by many small and unrelated random effects, are approximately normally distributed [43]. Now the readout signal voltage (for logic '0' or '1') obtained across the chip is normally distributed owing to the fact that there are several random effects and unrelated parameter variations in the read operation, which affect the readout voltage. Hence for our analysis the normal distribution was used to study the effect of parametric variations on readout voltages in the proposed multilevel FeRAM, i.e. we assume that the distribution of the readout voltages across the cells is *normal*. This was also backed up by the fact that all of the references in the literature show the charge distributions curves approximately normally distributed [21][37].

Since we had no actual data from a chip we needed a device to obtain normal distribution from simulation measurement for noise margin study between the different levels. With a chosen mean value of FeCap area, the area of the capacitor was varied by +-100% in steps of 5%. The readout voltages corresponding to V_H, V_M and V_L were plotted on the X-axis and the probability density function obtained from the intermediate values of the areas was plotted on the Y-axis [See Figure 5-5 (a)]. Separate distributions were obtained for the three nominal signals V_H, V_M, and V_L. Later these three plots were plotted on the same X and Y axes to give something like that shown in Figure 5-5 (b). The resulting cell signal distribution functions show available noise margins between the two adjacent distributions. The standard deviation for all the individual plots was later varied uniformly such that the adjacent distributions do not overlap. The value of σ at which they do not overlap appreciably was found to be +-15% of the mean.

Since all the corresponding PDF values were the same for all the three plots, the height of all the three distributions is the same. However, due to variations in the readout voltages their spreads were different and hence the area was different. To observe the relative spread and distribution, the areas of the curves were normalized by dividing the areas by the smallest value of the area (out of the three areas of the curves V_H, V_M and V_L). This technique formed the

basis of noise margin analysis for all further simulations, i.e. for any parameter variations distribution plots were obtained and noise margins were studied. To simplify and speed up the process of characterization, all of the various possible cell parameter variations were lumped into the cell capacitance (area). For example, sense amplifier offsets over a large cell array can be modeled as random errors that can be lumped with the cell area variations. Since C_{BL} and C_{FE} are probably the two dominant parameters that affect the readout voltages of all the three levels, the variations in the several parameters can be lumped into variations in the cell capacitance and BL capacitance. The various parameters were thus varied by +/-20% of their mean and their effect was observed on V_M, V_H and V_L.

To evaluate the effects of variations in cell area on the readout voltages, a series of simulations were run over a range of values of FeCap areas. Figure 5-6 shows the simulation results. It can be observed that with smaller bitline capacitances, the bitline voltage strengths increase. However, if the bitline capacitance is too small the noise margins between V_H, V_M and V_L becomes too small. Similarly, with increases in the cell area voltage levels increase as well. Again merely increasing the area does not give the best results. To better understand this, a second set of simulations was performed with constant area and varying bitline capacitance. It was concluded that there has to be an appropriate ratio between the bitline capacitance and ferroelectric capacitance to get the best cell signal results. This is shown by plotting the V_H, V_M and V_L curves for a range of C_{BL}/C_{FE} ratios. The spread in the cell signal distribution is a direct projection of the readout voltages values on the Y-axis in the Figure 5-6. So the objective of the design has to be to select the area of the capacitor where the slope of the curves of the readout voltages is the least. This is due to the fact that wherever (for a specific area) the slope is least there would be less scatter in the voltage distribution. In Figure 5-6 we see that we have the least slope where the area of the capacitor is around $3\mu m^2$. Hence this area should be selected for the design. However, since this is too large for the current 0.35μm technology (in which this model simulation was tried) the next smallest area of $1\mu m^2$ with least slope was selected.

(a) Bitline signals weighted by the normal distribution for one (V_H or V_M or V_L) signal value.

(b) V_H, V_M and V_L signal distributions plotted on same X-axis

(c) Optimized standard deviation δ for least overlap

(d) Normalization of areas to get resultant probability density function

Figure 5-5 Normal Distributions for the Readout Voltages

Figure 5-10 shows the distribution plots obtained from schematic simulations with appropriate C_{BL}/C_{FE} ratios chosen. FeCaps with different areas of 0.25 μm^2, 0.65 μm^2, 1.0 μm^2, 2.0 μm^2 and 3.0 μm^2 were simulated with C_{BL} of 200 fF, 250fF, 300fF, 300fF and 400fF respectively. Note that most of the distributions do not overlap for standard deviations of 10 to 20% of the mean. However, for the FeCap with an area of 3.0 μm^2 allows 25-30% variations. This can be explained from the curves in Figure 5-6. Since the slopes for all the three readout voltages are least for this mean area (3.0 μm^2) the corresponding distribution has the least spread. So the objective in this sort of simulations should be to obtain the least slope for the selected value of area and bitline capacitance. One interesting conclusion that can be drawn from these sets of distributions is that the standard deviation for all the ranges of area, which gives optimum results (no overlaps of the distribution), is around 15% of the mean. Figure 5-9 shows the 3D plot for the distribution and gives a rough idea about the noise margins for various values C_{BL} and cell area. Again it could be seen that proper ratio of C_{BL}/C_{FE} is necessary. If C_{BL} is too high (for a given cell area) noise margin between V_M-V_L reduces and if it is too low the noise margin between V_H-V_M reduces. A similar trend could be observed for variation in cell area for a particular C_{BL}.

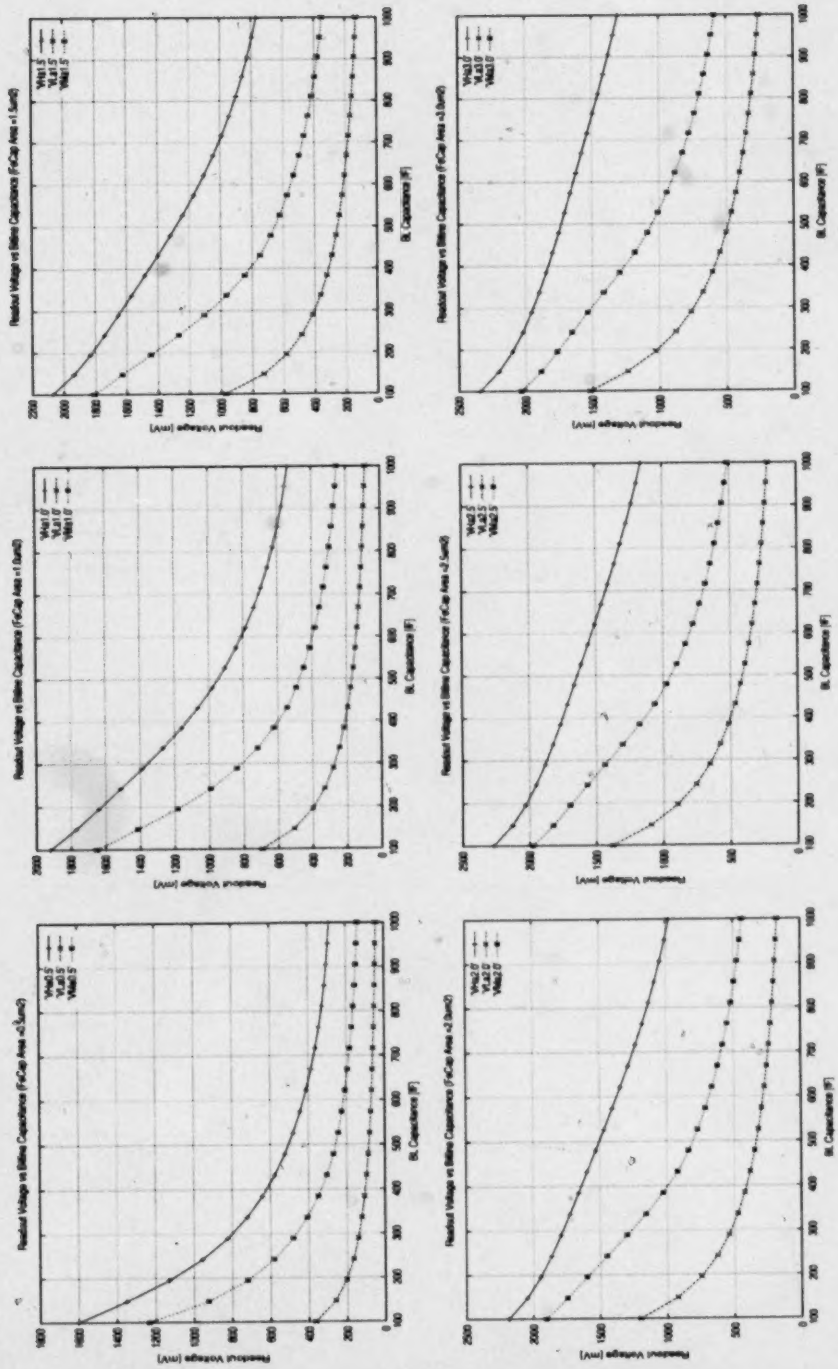


Figure 5-6 Variations in V_H, V_M and V_L with Variations in the FeCap area

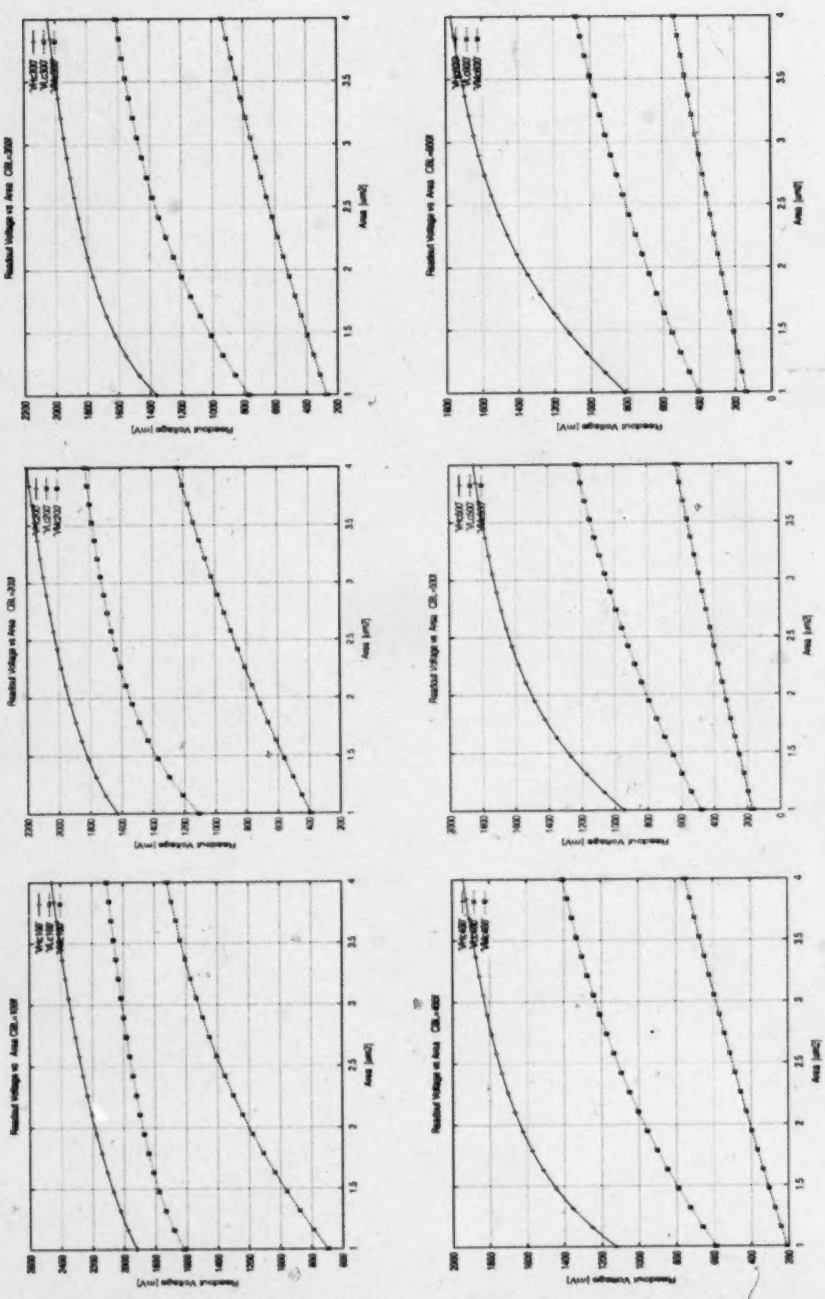


Figure 5-7 Variations in V_H, V_M and V_L with Variations in the Bitline Capacitance C_{BL}

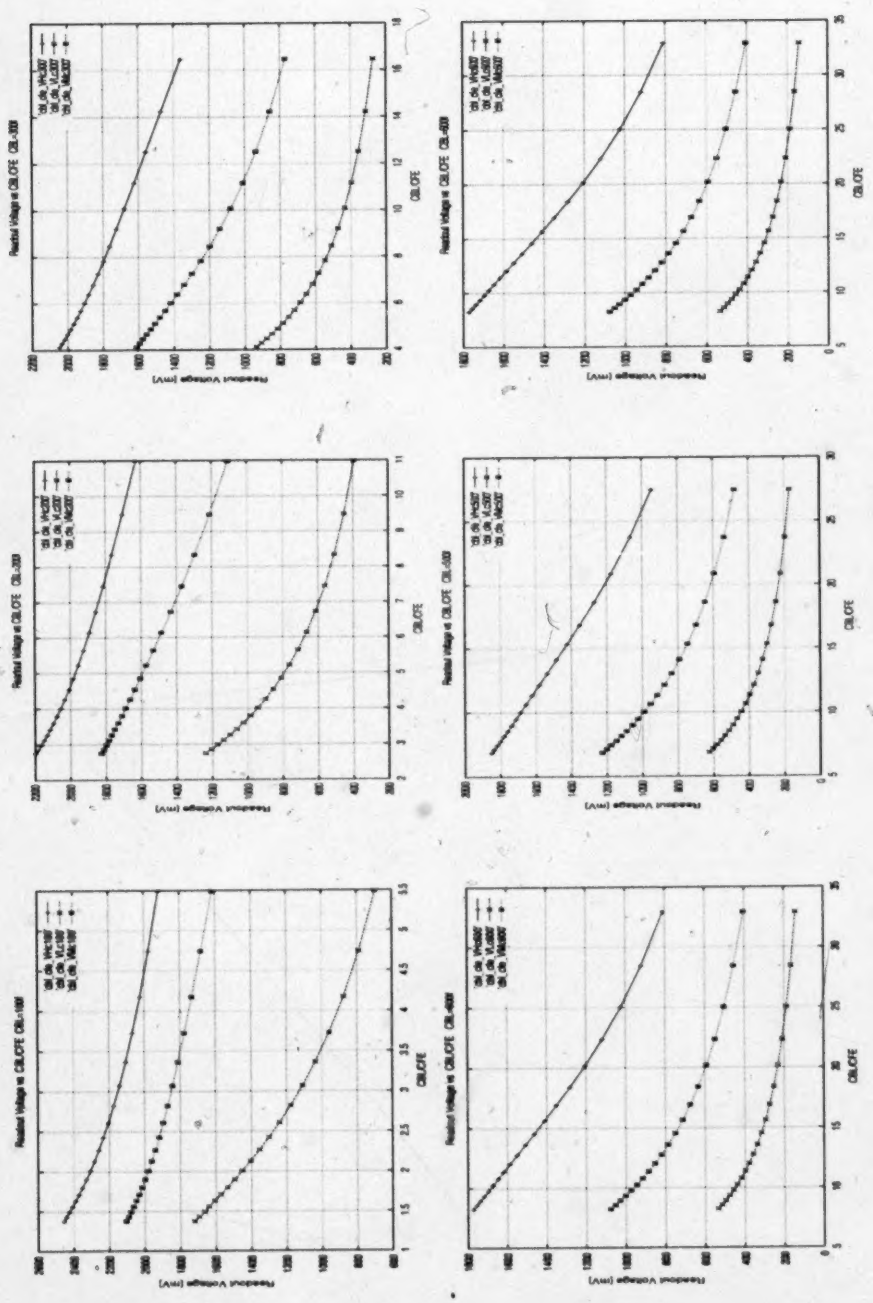


Figure 5-8 Variations in V_H, V_M and V_L with Variations in the C_{BL}/C_{FE} ratio

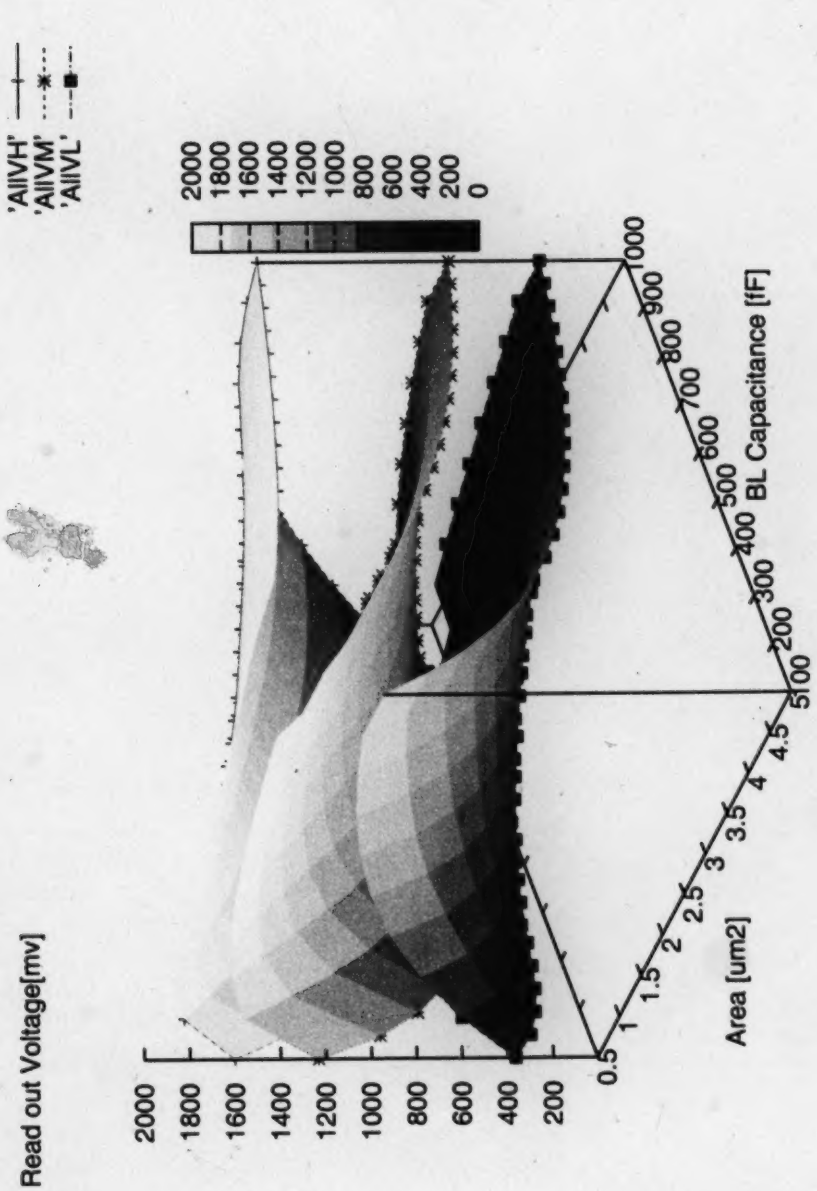


Figure 5-9 Readout Voltage vs. Cell Area and BL Capacitance

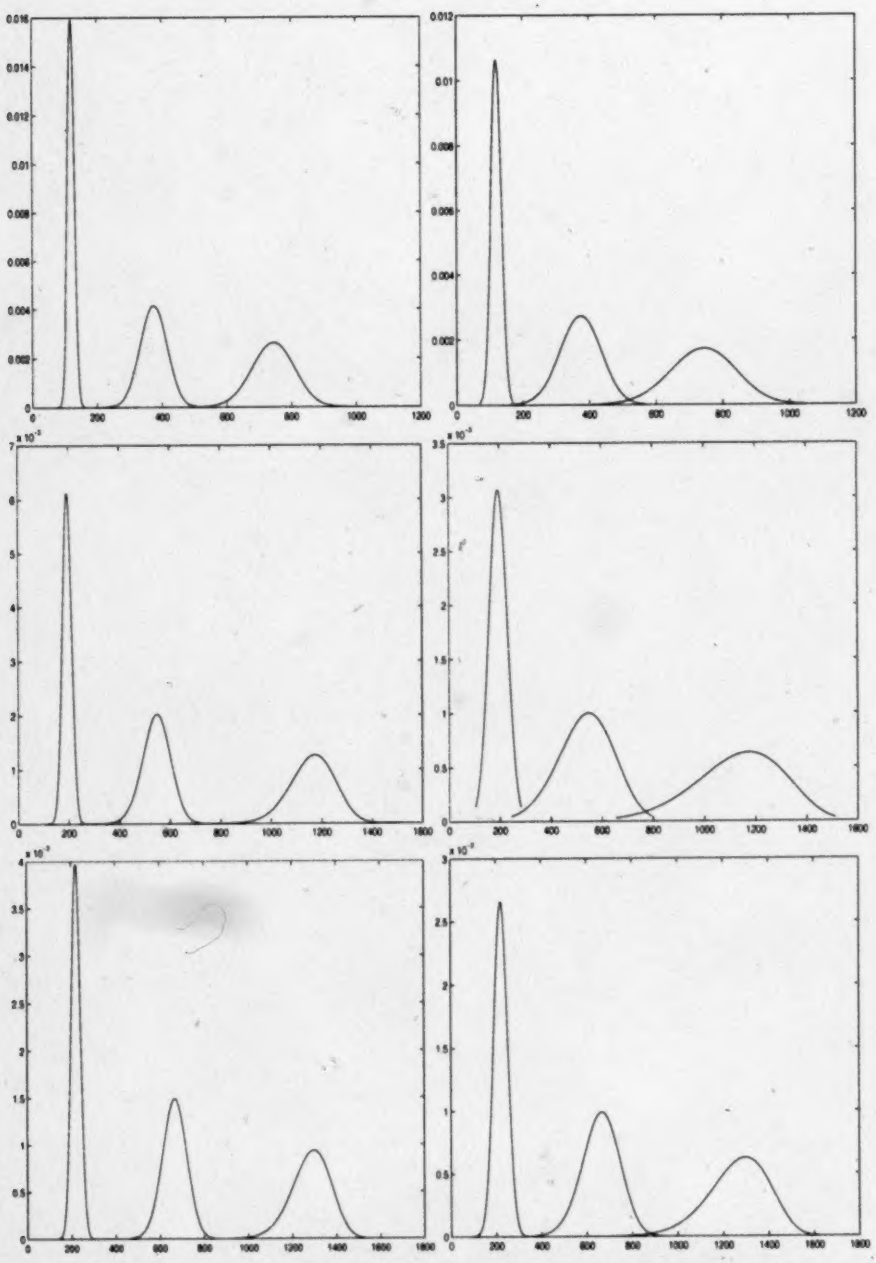
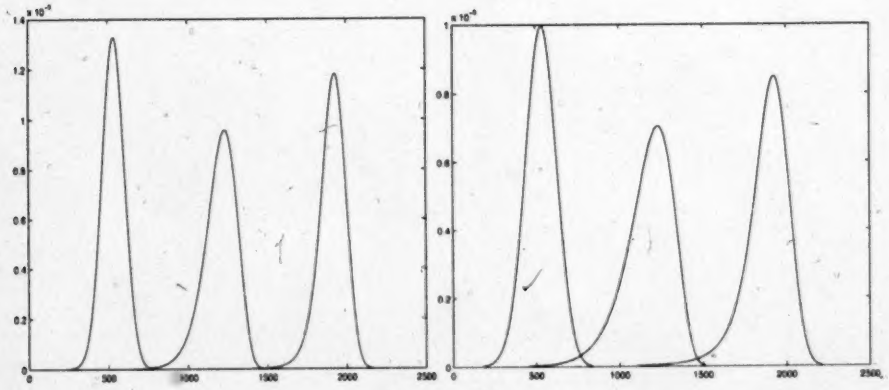
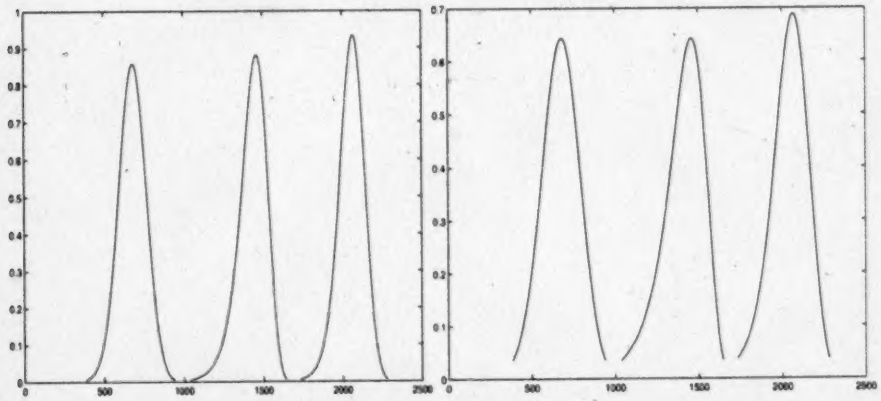


Figure 5-10 Effects of 10% (left) and 20% (right) Standard Deviations in Cell Areas of (Top) 0.25μm^2 cell (Middle) 0.65μm^2 cell (Bottom) 1μm^2 cell [X-Axis = Readout Voltage in mV, Y-Axis PDF]



(a) Cell Standard Deviations of 15% (left) and 20% (right)



(b) Cell Standard Deviations of 20% (left) and 25% (right)

Figure 5-11 Cell Signal Distributions for Cell Sizes (a) $2\mu m^2$ (b) $3\mu m^2$

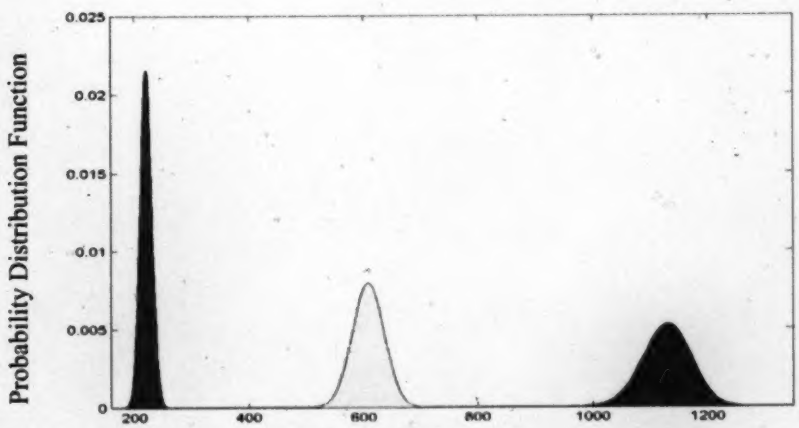
[X-Axis = Readout Voltage in mV, Y-Axis PDF]

5.3.3 Variations in V_M

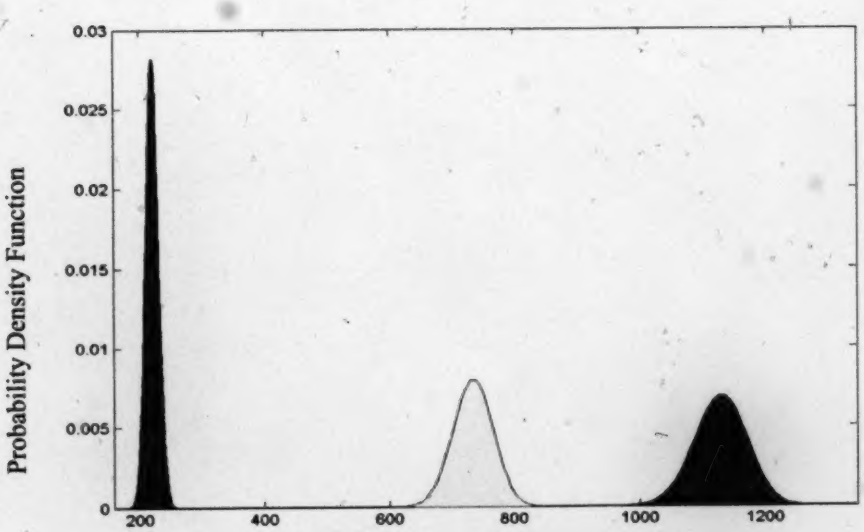
Although the V_M driver is designed for robust and stable operation, the output voltage might be susceptible to minor fluctuations during the dynamic operation of the chip. Also it is vulnerable to process mismatches and device ageing. Hence it is important to have a rough estimate of the impact of its variability and its impact on the reliability of the working of MLFeRAM. For a 10% variation in the cell area for the distribution of voltage levels, the V_M was varied by +/-5%, +/-10% and +/-15%, and the resulting cell signal distributions were constructed. It was concluded that roughly +/-15% variability in its (V_M) value is acceptable for reliable operation. The obtained distributions are shown in Figure 5-12.

Table V Noise Margins for Variations in V_M

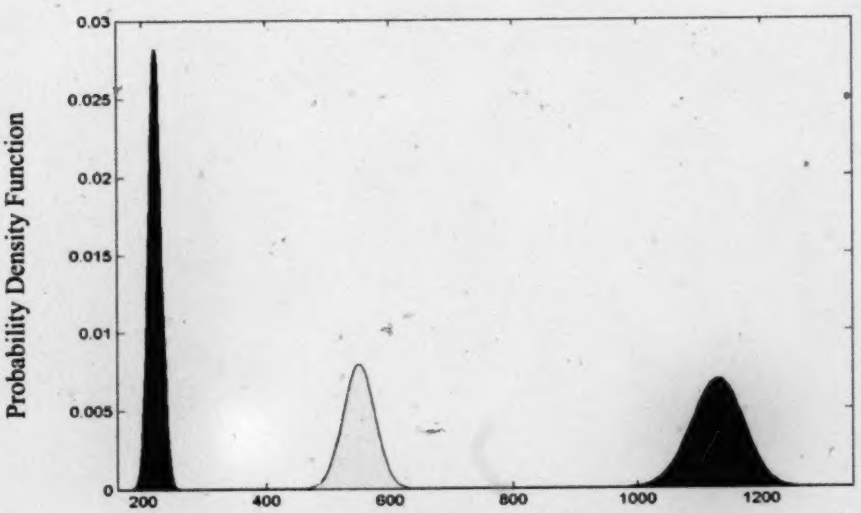
Std deviation (%)	N_{VH}-N_{VM} (mV)	N_{VM}-N_{VL} (mV)
-5(+5)	300(375)	300(170)
-10(+10)	220(450)	350(100)
-15(+15)	170(500)	420(10)



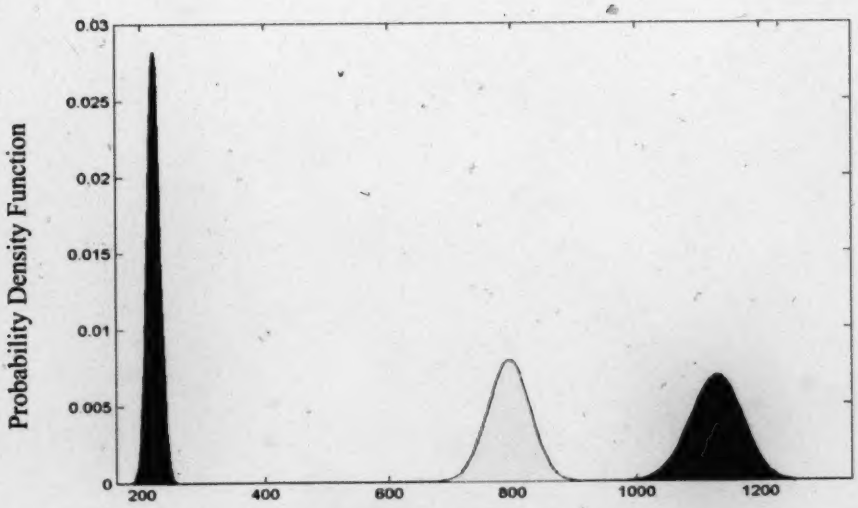
(a) –5 % of the mean V_M



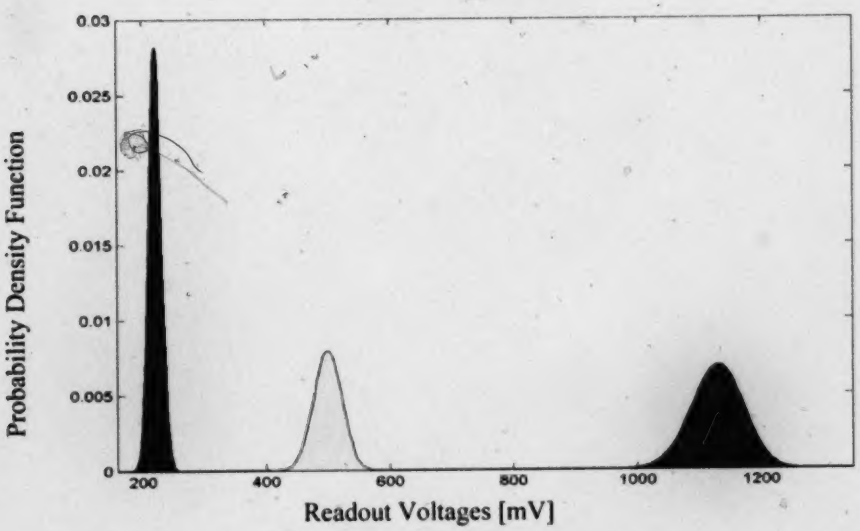
(b) +5 % of the mean V_M



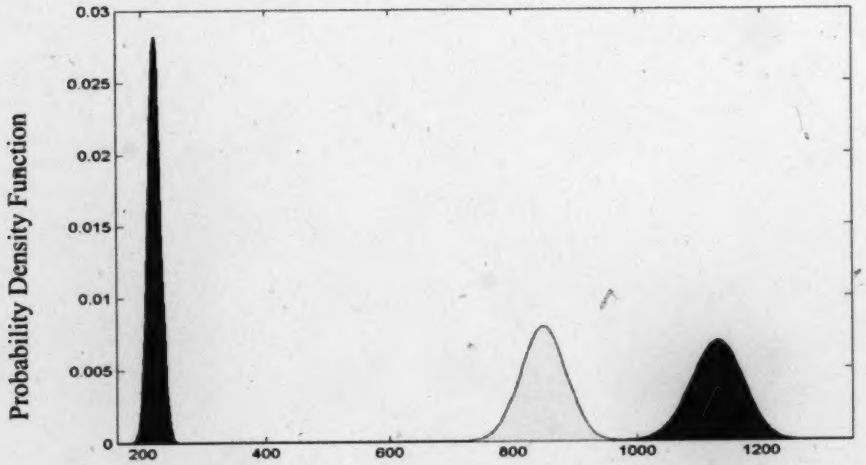
(c)-10 % of the mean V_M



(d) +10 % of the mean V_M



(e)-15 % of mean V_M

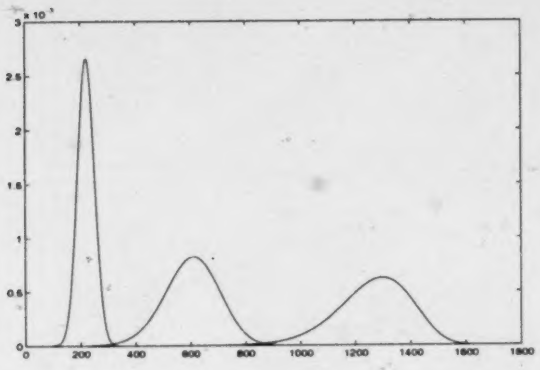


(f) +15 % of mean V_M

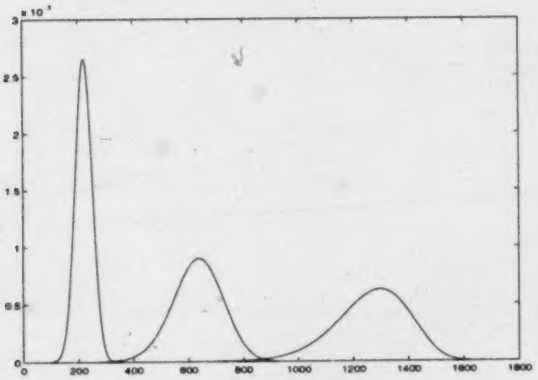
Figure 5-12 Distributions with Variation in V_M (a) Minus 5% (b) Plus 5% (c) Minus 10% (d) Plus 10% (e) Minus 15%(f) Plus 15% [X-Axis = Readout Voltage in mV, Y-Axis PDF]

5.3.4 Implementation Alternatives and Cell Signal Distributions

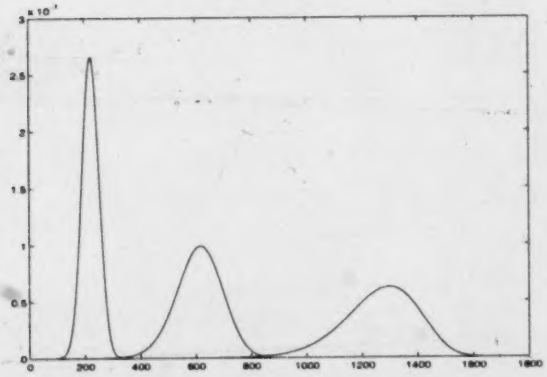
As discussed in Section 4.3.1 and shown in Figure 4-10, reaching a zero polarization state is very difficult and failure to reach zero polarization leads to wider distributions in the readout voltage of V_M. Figure 5-13 (a) shows the original distribution while (b) shows the distribution obtained with the help of subloops, which are obtained by giving a step input to the OPAMP. Part (c) shows the distribution obtained when a damping square wave is applied to the FeCap. The distributions obtained here are with σ of +-15% of the mean, area variation from 2.1 μm^2 to 3.2 μm^2 and bitline capacitance of 350fF with rest all the parameters same as mentioned in Section 5.3.1. According to the theory proposed in section 4.3.1 due to unrealizable depolarization the spread of V_M is large and should decrease if the sublooping technique is used. Accordingly a damped wave obtained from OPAMP and DAC was applied to depolarize the material.



(a)



(b)



(c)

Figure 5-13 Cell Signal Distributions with Implementation Alternatives: (a) Original (b) Damped Transient from the OPAMP (c) Damping Wave Generated from DAC

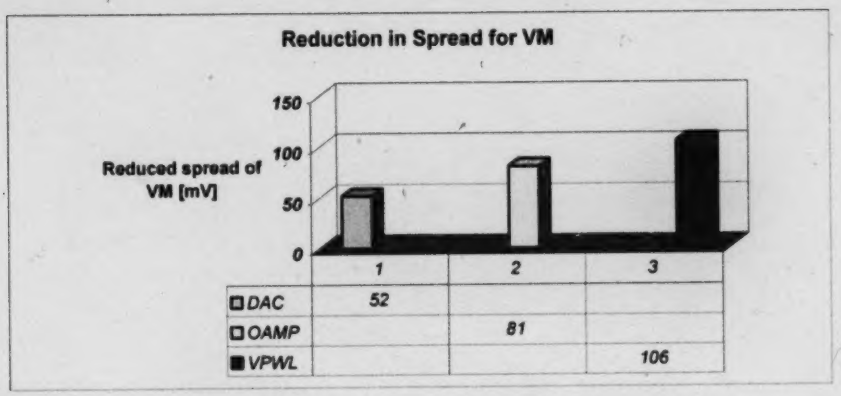


Figure 5-14 Reduction in the Spread of the Distribution of V_M with Various Techniques (DAC, OPAMP, VPWL[26])

Although the effect is less visible, tapering of the distribution of the middle level voltage is obtained in Figure 5-13 (b) and (c) compared to the original distribution in (a). Some of the reasons why one might expect an oscillating waveform to have advantage over simply writing a stable V_M are: 1) Difficulty in writing accurate V_M when so many device parameters can slightly vary. 2) A similar technique has proven to be effective when demagnetizing ferromagnets. To conclusively support this claim Figure 5-14 shows a quantitative comparison between the two techniques and the corresponding improvement in the distribution. VPWL was chosen because it generates an ideal damping square wave suitable for observing simulation results at the initial stage. Noise margins were calculated from the points where the deviation is +-2σ from the mean. This is justifiable because 95 % of the total values fall within +-2σ [See Section 5.2].

To reduce the effect due to the non-switching part (due to linear cell capacitance) of the cell signal, the cell signal distribution was observed after the WL and BL are driven low (this is done in pulse sensing technique [1]). This distribution in fact showed better noise margins between the levels. Figure 5-15 shows the comparison between the noise margins obtained with pulse sensing

[26] VPWL is an ideal piecewise linear voltage source in simulation used to obtain the damped wave.

(after pulse sensing) and step sensing (during pulse sensing) techniques. An advantage arises due to the fact that the readout voltages obtained for V_L is always zero causing the spread of this distribution to be less than 10mV. This gives us a larger (1800mV compared to 1200mV in step sensing) margin between the V_H and V_L distributions, which in turn allows us to place V_M closer to V_L than was possible in step sensing. This essentially increases the noise margins between V_H-V_M and V_M-V_L thus making the memory more tolerant to noise and parameter variations. However, the spreads of the distributions of individual V_H and V_L were unaffected. Following the similar approach as above here noise margins were calculated for deviations within +-2σ of the mean.

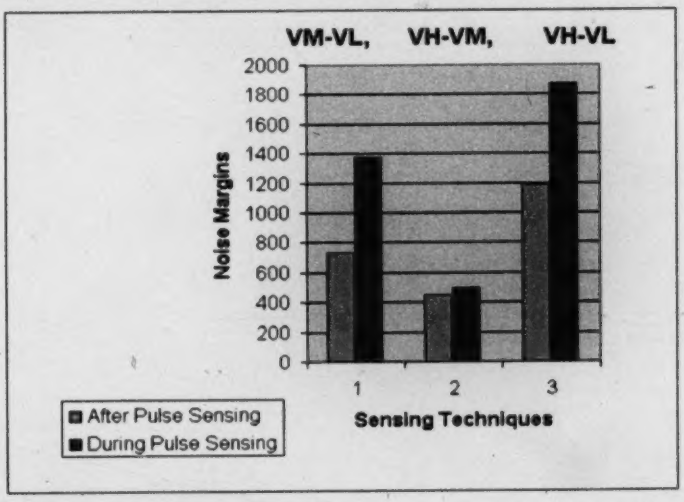
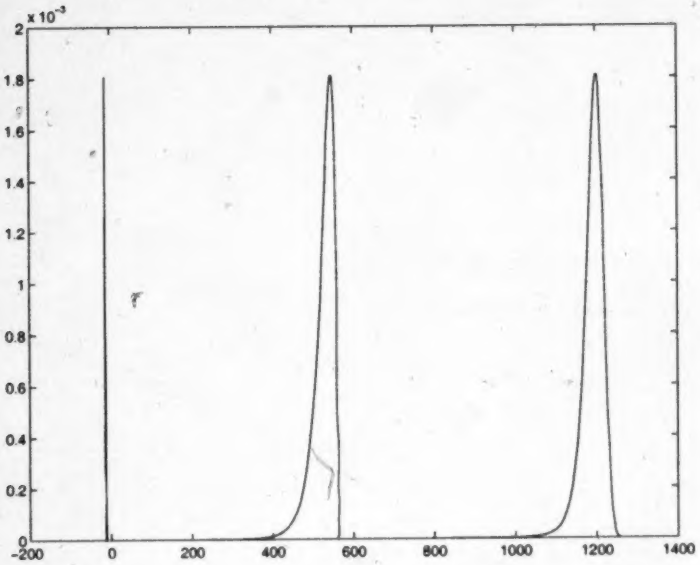
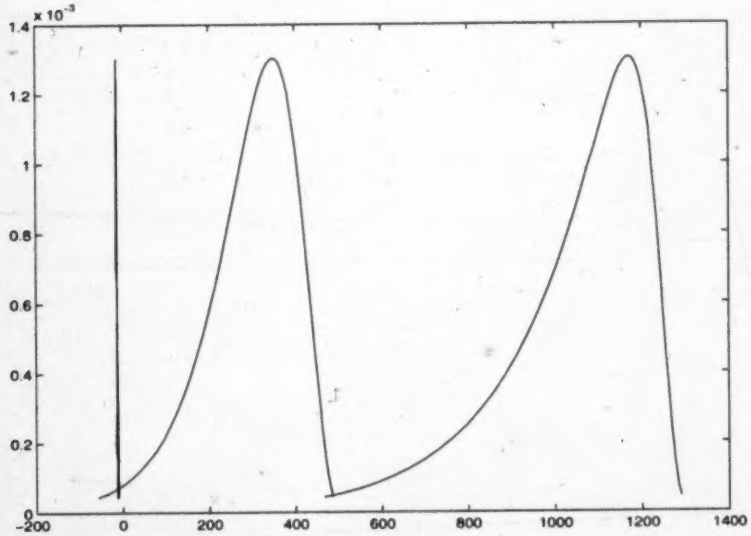


Figure 5-15 Noise Margin Comparisons for Alternative Sensing Techniques

Figure 5-16 shows the distribution plots for After-pulse sensing technique. A good improvement in the noise margins can be observed. However for higher values of the cell area variations long tails are observed. The tail is asymmetric and extends more towards negative side of the variation. This is attributed to the fact that there is a higher nonlinearity in the readout voltage corresponding to variation in area for smaller values of area for a give bitline capacitance. This could be observed in Figure 5-7. If by some means these tails could be truncated than huge improvements in the noise margins can be obtained.



(a)



(b)

Figure 5-16 Cell Signal Distributions With the After-pulse Sensing Technique: (a) 20% and (b) 40% Variation

5.3.5 Write Sequence Variations

Writing the middle level voltage V_M to depolarize the cell dielectric is a challenging task. We found it useful to consider some new PL and BL sequences. At the same time any new method must not hamper the writing of the two other cell voltages V_H and V_L. Out of the several approaches, two promising ones are shown in Figure 5-17. In both the methods it was found that the readout voltages show some (unwanted) dependency on the previous state of written voltage on the FeCap. This creates dependability and non-uniformity in the readout voltages. As shown in Table VI for both methods the readout voltage for '0' is different (200mV, 178mV and 176mV) if the previous value written is different (V_L, V_M and V_H respectively). Here 200mV is shown shaded because here there is a major discrepancy in the readout voltage. The same is true for a readout voltage of 1.5V when the previous value written is 3.3V. Though these deviations due to history dependency is just around 10-30 mV, in real cells this unwanted history effect might be worse due to parametric variations, power supply variations and process mismatch.

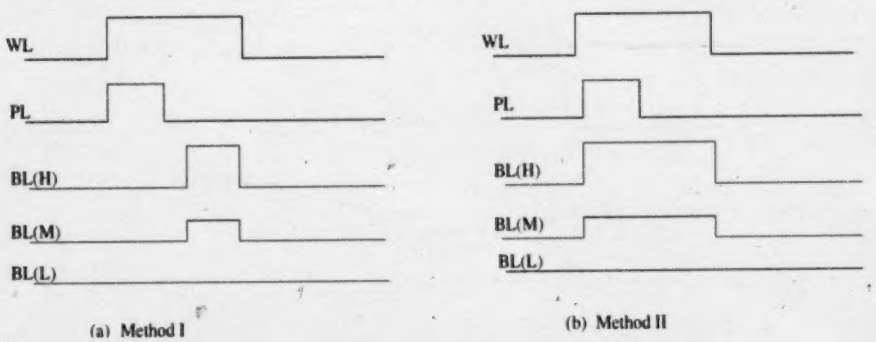


Figure 5-17 Pulsing Techniques: (a) Type I, (b) Type II

Table VI Readout Voltages (V_R) for Type I and Type II Pulsing Sequence

Type I				Type II		
V1	V2	V_R		V1	V2	V_R
0	0	200		0	0	200
0	1.5	562		0	1.5	595
0	3.3	944		0	3.3	949
1.5	0	178		1.5	0	177
1.5	1.5	563		1.5	1.5	595
1.5	3.3	943		1.5	3.3	949
3.3	0	176		3.3	0	176
3.3	1.5	555		3.3	1.5	584
3.3	3.3	944		3.3	3.3	949

This undesirable history effect was studied in more detail in simulation of up to three previously written voltages. Table VII shows the waveform voltages for sequence of three previous states (three previous voltages). V1 is the third-last

Table VII Readout Voltages after Writing a Sequence of Three Voltages

	V1	V2	V3	V_R (mV)	V1	V2	V3	V_R (mV)	V1	V2	V3	V_R (mV)
	0	0	0	204.09	1.5	0	0	180	3.3	0	0	179
	0	0	1.5	596.94	1.5	0	1.5	596.94	3.3	0	1.5	586
	0	0	3.3	952.63	1.5	0	3.3	952.63	3.3	0	3.3	952
	0	1.5	0	180.64	1.5	1.5	0	180.64	3.3	1.5	0	179
	0	1.5	1.5	596.83	1.5	1.5	1.5	596.70	3.3	1.5	1.5	586
	0	1.5	3.3	952.65	1.5	1.5	3.3	952.67	3.3	1.5	3.3	952
	0	3.3	0	179.13	1.5	3.3	0	179	3.3	3.3	0	179
	0	3.3	1.5	586.21	1.5	3.3	1.5	586.22	3.3	3.3	1.5	586
	0	3.3	3.3	952.69	1.5	3.3	3.3	952.68	3.3	3.3	3.3	952

state/voltage written, V2 is the second-last state written and V1 is the last voltage written before doing any read. This dependency study was important to perform

because in multilevel write operations the previous state is unknown and a new state has to be written. Hence it is desirable to go to a known state and from there go to the targeted state. The shaded sections in Table VII as well show the major inconsistencies in the readout voltage. We can observe though that in the extreme right set of V1, V2, V3 in the table there are no discrepancies provided V1 stays 3.3 V no matter what state represents V2 the readout voltages of V3 are consistent. This was ensured by writing 3.3 V into the FeCap at the start of every simulation cycle. One possible reason is that this could be due to an initialization problem in the FeCap model. But if it were due to this the same effect could have been observed for two-level simulation. Also it is unclear why the model has to be initialized to the 3.3 V state and not others. Nothing substantial has been concluded out of this. A test chip characterization might be needed to see if a similar effect is observed. However this effect was important to note down and was taken care of at the initial stage because otherwise this might have lead to unreasoned discrepancies in the noise margin analysis at the later stages of the design.

Chapter 6

Conclusions

6.1 Summary

Ferroelectric memories have already shown good potential to be used as next generation non-volatile memories. They have many advantages such as high endurance, faster writes, low-voltage and low-power operation, CMOS process compatibility, etc. in comparison to other contemporary non-volatile memories such as Flash and EEPROM. They have been used in quite a wide range of applications including RFID circuits, embedded memory and stand-alone memory. However, available FeRAM densities have been evidently limited (commercially available in sizes no greater than 1Mb) due to the larger area occupied by the cells and the older 0.35-μm process available from the fabrication line used by both Fujitsu and Ramtron.

The main focus of this work has been to evaluate a new direction towards increasing the density of FeRAMs. Similar to the concept of multilevel designs in DRAM and Flash memory, the MLFeRAM uses more than two signal levels to be stored in a cell. Here we will first evaluate some of the potential advantages and

necessities for this idea. As the processing technology for ferroelectric materials continues to mature with time it may become possible to store multiple valued polarization on a single capacitor [40]. To achieve this in practice, precise characterization of such a multilevel storage cell is needed. The multilevel idea is also motivated by the fact that we want to reduce the cost per bit. Scaling the technology does help to increase densities of chip. But this is achieved with the extra amount of cost needed for setting up a new fabrication line for new dimensions. It might be more challenging and costlier to increase memory density by reducing the feature size. In a review [38] on the prospects of emerging new memory technologies, further scaling of FeRAM has been questioned unless new technologies like 3-D capacitors and ultra thin-film technologies can be developed. In that respect, further increases of density might be achieved in part by applying some concept like MLFeRAM. The improvement in density could be appreciated by the amount of gain we might have. For example a 256 Mb chip will now be 384 Mb without any need for further scaling. To add to this, a programmability feature in the chip could be added so that if for some reason the cells do not work for three levels, the chip could still be used as a two-level conventional FeRAM. Hence multilevel operation does not affect the yield in terms of two-level memory.

The idea of storing multilevel polarizations on a single cell was simulated assuming a 0.35-μm technology. The ferroelectric capacitor model developed in [3], which is based on *Preisach* theory [4], was integrated into the Cadence environment for capturing the behavior of FeCap. A 1T1C structure was used as a memory cell and a range of simulations were run for varying sizes of ferroelectric capacitors and bitline capacitances. The readout voltage distributions (based on a Normal distribution) for all three levels of the cell were plotted. It was concluded that the idea indeed works in simulation and should be realizable for standard deviations of up to 15 percent of the mean cell area value. As mentioned in [16], it was observed that there is no significant reduction in switched polarization as the capacitor sizes are varied in the range from $10\mu m^2$ to $1\mu m^2$. Hence the variation in the readout voltage should mostly be attributed to the non-switching part of

polarization. This implies that the idea is not much affected even for smaller cell sizes unless there is a potential difference in the remnant polarizations of the cell.

To verify the idea a cell signal voltage distribution graph was obtained from [21] and is presented graphically in

Figure 6-1.

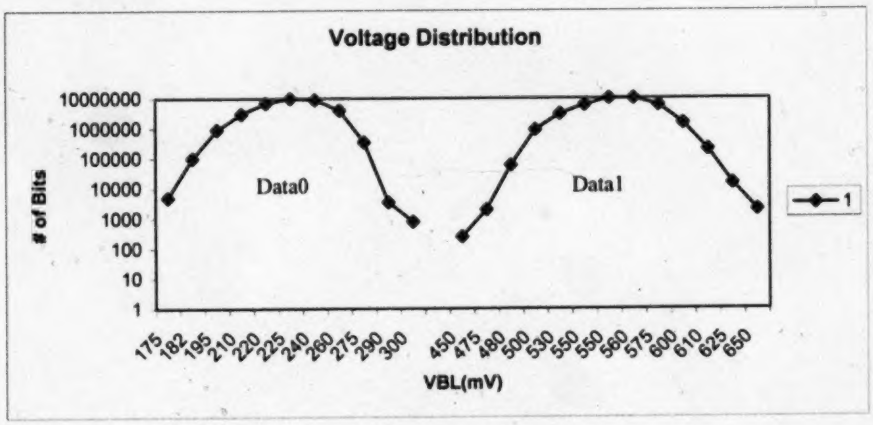


Figure 6-1 Voltage Distribution due to Overall Parametric Variations [21] Log-scale

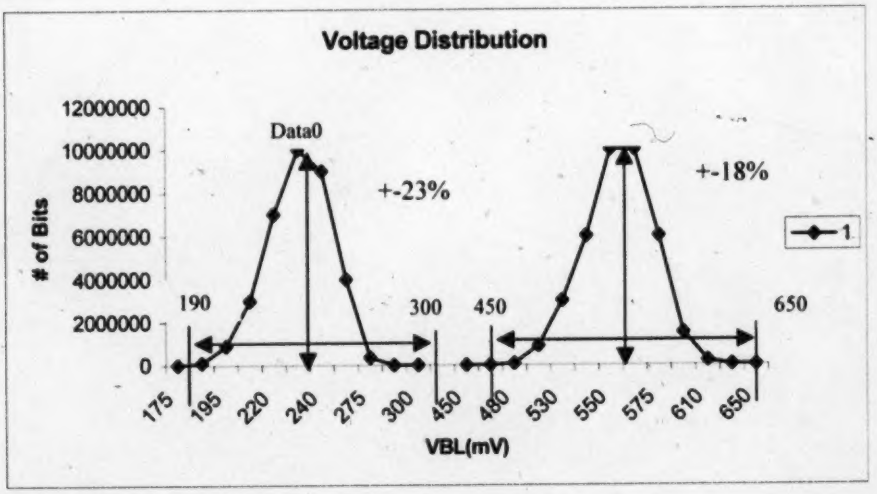


Figure 6-2 Voltage Distribution due to Overall Parametric Variations [21] Linear-scale

This distribution was on a log scale in the original source. To do a comparative study with the plots obtained in our simulations this voltage distribution was re-plotted on a linear scale as shown in

Figure 6-2. The Data 0 and Data 1 have standard deviations of about +/-25% and +/-18% respectively, about the mean. In MLDRAMs a noise margin of about 100mV is required between two levels out of which roughly 40mV margin is kept for sense amplifier offsets and roughly 60mV for external noise and voltage variations due to inevitable device parameter variations. A comparable minimum acceptable noise margin is also anticipated in FeRAMs. We achieved (in simulations) this minimum amount of noise margins between the levels and also achieved 25% variation in the voltage distribution. In our simulations parameter scatter was modeled by varying the cell area. A cell area variation of 7-8% of the mean (0.25 μm^2) achieves resultant 25% variation in readout voltage. To provide a concluding (and comparative results to that cited in [21] and above Figure 4-5) result we took a pessimistic approach. We took a cell area of 0.25μm^2

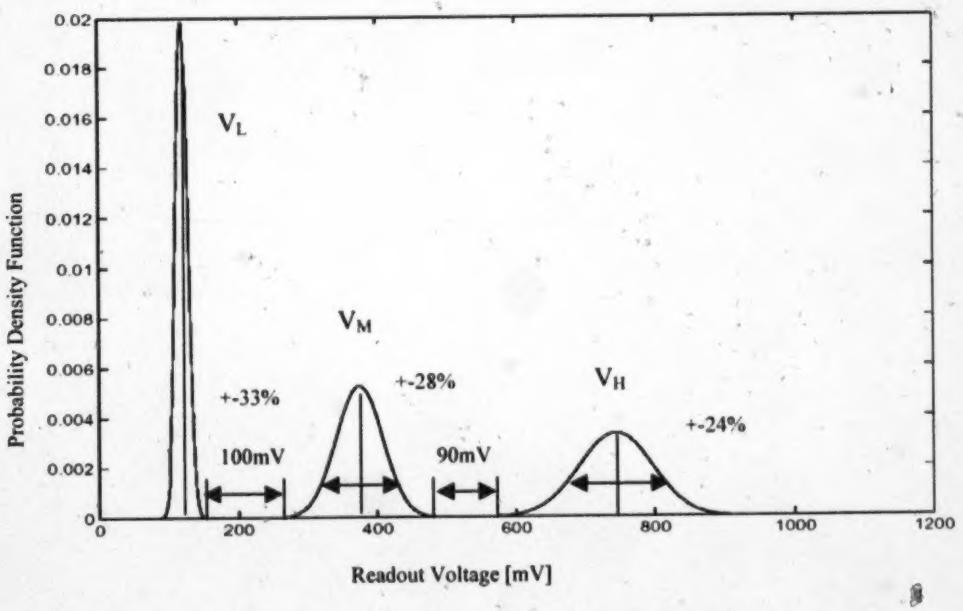


Figure 6-3 Voltage Distribution for a Three-level FeRAM with Variation in Cell Area

instead of 0.44 μm^2 taken in [21] and obtained readout voltages variations of +/-24%, +/-28% and +/-33% which are quite high compared to variations of +/-18% and +/-23%, and yet achieved good noise margins. Now these noise margins can be further increased by ~50mV by using various techniques (damping wave and after pulse sensing). It is important to note here that the above-mentioned distribution was only possible and is dependent on the charge retention (remnant polarization Pr) of the material. We have assumed the characteristics of PZT and its compounds, which has remnant polarization of around 20-25 $\mu C/cm^2$ [45]. On the other hand it might be difficult to obtain such[27] distribution with the materials like SBT employed in Hynix Ferroelectric process and Matsushita ferroelectric process, which have switched charge of around 15 $\mu C/cm^2$ and 8 $\mu C/cm^2$ [46] respectively.

To summarize a range of implementation ideas was explored. A new method of read and write sequence was proposed. A voltage driver based on a traditional two-stage differential amplifier OPAMP was designed for writing the middle level voltage. To better depolarize the material to the zero state, a subloop technique based on applying a damped oscillating wave to the FeCap was used. A 5-bit digital-to-analog converter was also designed for generating a square-shaped damped wave. This DAC is based on a conventional high-speed current steering technique, which has a higher slew rate compared to conventional DACs, which is important in our application to keep the write speed as fast as possible. The distribution plots were observed to show some improvement with the application of the proposed techniques. The readout voltages for the *after pulse sensing* scheme show excellent distribution with good amount of robustness against variation in cell area. This method should be retained and further tried out for any actual MLFeRAM distribution measurement.

In a conventional FeRAM some memory cells are used as reference cells. Each column or row might have one reference cell for generating the reference voltage. Due to intrinsic similarities in the material, the reference cell tracks the

---

[27] Distribution which enables higher noise margins for the readout voltages.

variations in the main memory cells. However, due to limited endurance characteristics and the much greater number of accesses, the reference cell is more prone to fatigue and degradation than the data-storing cells. Generating an accurate reference is critical for reliable multilevel cell operation and may unfavorably affect its functionality. To overcome this issue an auto-calibrated reference voltage generation scheme is proposed. This scheme uses charge-sharing technique similar to that used in MLDRAM. Prior to using the FeRAM, a range of the reference voltages is swept through the multilevel cell. Based on the results of this a bath-tub curve is plotted and appropriate reference voltages are chosen between the available noise margins (V_{ref1} between V_H-V_M and V_{ref2} between V_M-V_L). This information is stored in the flip-flop for further operations.

It is our belief that the MLFeRAM idea presented in this work could be a reasonable new direction for increasing FeRAM densities. Our simulation results suggest that the idea is promising enough to be tried out on a test chip.

The contributions of this work include:

1) Feasibility analysis of a three-level FeRAM. (Cell design, Noise margin improvement)

2) Design of an OPAMP for writing the middle level signal voltage.

3) Design of a DAC-based bitline driver for generating subloops in the hysteresis.

4) Schematic-level simulation of multilevel FeRAM.

5) Auto-calibrated reference generation technique.

6) A simulation-based study of the effects of device parameter variations on the cell signal distributions. (Based on a Gaussian distribution model)

6.2 Challenges

MLFeRAM faces multiple technical challenges to its use in commercial or industrial designs. In the following we will consider some of the major challenges associated with the new idea.

The read and write times have shown to be increased by around 20-30ns in simulations. This is due to the required modifications in the pulsing sequence of

PL and BL. If the subloop technique is employed for writing V_M, the access time increases by 80-90ns in the case when the damping waveform is obtained with a step input applied to OPAMP and by ~60ns if a DAC is employed. Note that since the read is destructive in a FeRAM, the restoration after each read also adds to the basic read time and could be large if sublooping is employed.

Since we are storing more than two levels on a single cell, extra decoding logic is required for the read operation (two cells have to be read at a time). This adds circuit complexity, area and an access time penalty. There is area overhead due to the sense amplifiers needed at both the ends of the bitlines. As well there will be an increase in area because of the reference generation logic required. The extra area and complexity of this decoding logic might not be much of a burden as the size of the memory chip increases; however, longer accesses times will definitely be an additional liability.

Reliability issues might severely hinder the acceptance of MLFeRAM. Fatigue is a major concern since it scales down the readout voltages. This will be due to the higher number of read and write accesses needed because of the fact that now one cell is storing a greater number of bits. As well, the changes in the write operations (recall that now we have to go to V_H state before writing any new state) almost doubles up the cycling of zero to one. The sublooping technique would increase the amount of fatigue in each cell because it increases the number of polarization reversals. Since in sublooping we do not do full switching (switching at saturation voltages) it is not well understood how the sublooping technique will fatigue the cell. On the other hand, the advantage of accessing a cell higher number of times will imply a higher imprint and relaxation immunity. Since a higher number of accesses to a cell means a higher probability of switching from one state to other (suitable for imprint) and a lesser amount of time that the capacitor is left unaccessed (suitable for relaxation[28]). A characterization on a FeRAM test chip may be required to justify any claims.

---

[28] Relaxation is a major concern in conventional FeRAM since it dynamically changes readout voltages and reduces polarizations to more than 25 % [14].

These issues could be handled by bringing some additional improvements in the process technology in near future.

6.3 Suggestions and Future Work

Since we did not have access to a real ferroelectric memory process, it was not possible to fabricate a chip with the proposed design. Currently only a few companies, including TI, Radiant Technologies, Matsushita and Fujitsu, have access to the FeRAM process. We believe however, that the results of our simulation study are promising enough to justify further work on multilevel FeRAM in a prototype chip.

From an analysis point of view we have the following suggestions that would be encouraging for multilevel FeRAM realization:

(I) A tighter control on process parameters is desired leading to smaller cell variations. This will be possible in the near future, as the ferroelectric process becomes more mature. For the present processes cell area variation limit of up to 10% or less is desirable (assuming that all the parametric variations are lumped as area variations). Readout voltage variations of less than 25-30% of the mean are desirable.

(II) Higher remnant polarization materials should be employed to achieve higher signal margins with reduced cell area. PZT is composed of $PbZrO_3$ and $PbTiO_3$. Traditionally a P_r of 18-30 $\mu C/cm^2$ is typical for 40/60 PZT. Higher polarization values of $81\mu C/cm^2$ can be obtained out of pure $PbTiO_3$ [3]. With this multilevel polarizations would be possible.

(III) The Fatigue being an electrically reversible phenomena could be addressed in two possible ways. (1) Adequate improvements must be made in the material processing of the basic ferroelectric composition to make it fatigue resistant. Recently addition of Niobium into PZT (PZTN) and SBT (SBTN) has shown [50] a strong improvement in the fatigue properties with PZTN showing little or no fatigue. (2) It has been observed [14] that application of a periodic polling pulse of 5 V (every 10^9 cycles) to a capacitor cycled at 2.5 V restores remnant

polarizations to nearly original levels. This shall be occurring less frequently (of the order of tens of seconds) compared to the DRAM refresh cycles.

(IV) The properties of aged capacitors can be recovered (remnant polarizations to original levels) by cycling the capacitor at 5 V for 10^4 cycles (@100KHz) [14].

(V) Niche applications (requiring high density, low speed and higher writes) could be targeted with multilevel FeRAMs that would provide a route for initial mass production to reduce cost per bit.

The following suitable topics need to be worked out in direction of MLFeRAM implementations:

(I) Sensing Scheme and Sense Amplifier:

In MLFeRAM the readout voltages have narrower noise margins between different levels than conventional FeRAM. In our work the cell signal distributions were measured in simulations for two widely used sensing schemes, i.e. step and pulse sensing. Additional sensing schemes should be investigated that will be fast as well as less prone to variations in the architecture of the design and the C_{BL}/C_{FE} ratio. One positive step achieved in this direction is proposed in [20] in which the charge stored on the capacitor is read rather than the bitline voltages. However this scheme has a greater amount of complexity and a larger SA design. Equal thought has to be given to the design of an appropriate sense amplifier, which would have less common mode voltage, higher sensitivity and good tradeoff between speed and area.

(II) Intermediate Voltage Driver:

One bitline driver designed in this work was based on a simple two-stage differential operational amplifier, which has its own limitations such as temperature dependability, limited PSRR and CMRR. Higher stability and accuracy of the middle level voltage driver is desired for achieving the zero polarization corresponding to the intermediate state. The driver should produce shrinking subloops so as to make readout voltages for the intermediate state, for the cells across the chip, less dependent on the individual cell variations. The driver should be able to depolarize the cell dielectrics to a state, which will give an accurate intermediate readout voltage that is exactly midway between high and

low. Another big challenge for this driver would be to track the variations in the cell characteristics because, as the cells fatigue, V_H and V_L will change and a new V_M will have to be written to maintain the best possible noise margins between the levels. This might involve some sort of auto-calibration in the driver, which is dependent on the cell characteristics. The bitline driver must also be able to successfully drive a variable number of cells. The number of cells requiring V_M will change depending on the written data.

(III) Ferroelectric Capacitor Modeling:

Accurately modeling the behavior of a FeCap seems to be a continuing challenge faced by FeRAM circuit designers, as most of the work these days is based on circuit simulator tools. The model employed in this work uses the arc-tan function, which is best suited to the physical properties of SBT. Other available models, which closely model the characteristics of other materials, could be tried out for the study of the distributions. Most of the Ferroelectric Capacitor (FeCap) models available to date are based on behavioral modeling: i.e. they model the ideal hysteresis loop characteristics of the material. They are based on various assumptions like symmetric hysteresis loops with respect to origin and frequency and temperature-independent hysteresis loops. Also, since they are designed for a two-level memory cell, they are mostly concerned about the accuracy of outer loops produced by large signal operation and less about the concentric loops closer to the origin (i.e. the near-depolarized operating region). However, for a multilevel cell, subloops near the origin are very critical for correct operation. More work is required to assess the available models and to modify them as necessary to ensure that they match with the real FeCap characteristics and hence enable multi-level operation. It would be really useful to have these models accurately represent non-ideal phenomena like fatigue, retention and imprint.

Bibliography

[1] Ali Sheikholeslami and P. Glenn Gulak, "A survey of circuit innovations in ferroelectric random-access memories", *Proceedings of the IEEE*, May 2000, Vol. 88, No. 3, pp. 667-689.

[2] Ali Sheikholeslami, "Transient modeling of ferroelectric capacitors for nonvolatile memories", M.A.Sc. Thesis, University of Toronto, ON, Canada 1994.

[3] Juergen Thomas Rickes, "Circuit Design of Gigabit Density Ferroelectric Random Access Memories", PhD Dissertation, Rheinland-Westphalia Technical University, Germany, December 2002.

[4] Andrei T. Bartic, Dirk J. Wouters, Herman E. Maes, Juergen T. Rickes and Rainer M. Waser, "Preisach model for the simulation of ferroelectric capacitors", *Journal Of Applied Physics*, March 2001, Vol. 89, No. 6, pp. 3420-3425.

[5] Daisaburo Takashima, Yoshiaki Takeuchi, Tadashi Miyakawa, Yasuo Itoh, Ryu Ogiwara, Masahiro Kamoshida, Katsuhiko Hoya, Sumiko Mano Doumae, Tohru Ozaki, Hiroyuki Kanaya, Koji Yamakawa, Iwao Kunishima, and Yukihito Oowaki, "A 76-mm^2 8-Mb Chain Ferroelectric Memory", *IEEE Journal of Solid-State Circuits*, November 2001 Vol. 36, No. 11, pp. 1713 – 1720.

[6] P.Zurcher, R.E.Jones, P.Y.Chu, D.J Taylor, B.E.White, S.Zafar, Bo Jiang, Yeong-Jyh Tom Lii, and S.J.Gillespie, "Ferroelectric Nonvolatile Memory Technology: Applications and Integration Challenges", *IEEE Transactions on Components, Packaging, and Manufacturing Technology—Part A*, June 1997, Vol. 20, No. 2, pp. 175-181.

[7] Daisaburo Takashima and Iwao Kunishima, "High-Density Chain Ferroelectric Random Access Memory (Chain FRAM)", *IEEE Journal of Solid-State Circuits*, May 1998, Vol. 33, No. 5, pp. 787-792.

[8] Ali Sheikholeslami and P. Glenn Gulak, "A survey of behavioral modeling of ferroelectric capacitors", *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, July 1997, Vol. 44, No. 4, pp. 917-924.

[9] Daisaburo Takashima, Susumu Shuto, Iwao Kunishima, Hiroyuki Takenaka, Yukihito Oowaki, and Shin-ichi Tanaka, "Sub-40-ns Chain FRAM Architecture with 7-ns Cell-Plate-Line Drive", *IEEE Journal of Solid-State Circuits*, November 1999, Vol. 34, No. 11, pp. 1557-1563.

[10] Y. J. Song, N.W. Jang, D. J. Jung, H. H. Kim, H. J. Joo, S. Y. Lee, K.M. Lee, S. H. Joo, S. O. Park and K. Kim, "Integration and Electrical Properties of Novel Ferroelectric Capacitors for 0.25μm 1 Transistor 1 Capacitor Ferroelectric Random Access Memory (1T1C FRAM)", *Japanese Journal of Applied Physics*, April 2002, Vol. 41, Part 1, No. 4B, pp. 2635–2638.

[11] H.Koike, T.Otsuki, T.Kimura, M.Fukuma, Y.Hayashi, Y.Maejima, K.Amantuma, N.Tanabe, T.Masuki, S.Saito, T.Takeuchi, S.Kobayashi, T.Kunio, T.Hase, Y.Miyasaka, N.Shohata and M.Takada, "A 60-ns 1-Mb nonvolatile ferroelectric memory with a nondriven cell plate line write/read scheme", *IEEE Journal of Solid-State Circuits*, November 1996, Vol. 31, No.11, pp. 1625 – 1634.

[12] T. Mikolajick, C. Dehm, W. Hartner, I. Kasko, M. J. Kastner, N. Nagel, M. Moert and C. Mazure, "FeRAM technology for high density applications", *Microelectronics Reliability*, 2001, Vol.41, No.7, pp.947-950.

[13] Y. Eslami, A. Sheikholeslami, S. Masui, T. Endo and S. Kawashima, "A Differential-Capacitance read Scheme for FeRAMs", *2002 IEEE Symposium On VLSI Circuits Digest of Technical Paper*, pp. 298-231.

[14] S.W.Wood, "Ferroelectric Memory Design", M.A.Sc Thesis, University of Toronto, ON, Canada, 1992.

[15] J.W.K. Siu, Y. Eslami, A. Sheikholeslami, P.G. Gulak, T. Endo and S. Kawashima, "A current-based reference-generation scheme for 1T-1C

ferroelectric random-access memories", *IEEE Journal of Solid-State Circuits*, March 2003, Vol. 38, No.3, pp. 541 – 549.

[16] Hugh P. McAdams, Randy Acklin, Terry Blake, Xiao-Hong Du, Jarrod Eliason, John Fong, William F. Kraus, David Liu, Sudhir Madan, Ted Moise, Sreedhar Natarajan, Ning Qian, Yunchen Qiu, Keith A. Remack, John Rodriguez, John Roscher, Anand Seshadri, and Scott R. Summerfelt, "A 64-Mb Embedded FRAM Utilizing a 130-nm 5LM Cu/FSG Logic Process", *IEEE Journal of Solid-State Circuits*, April 2004, Vol. 39, No. 4, pp. 667-677.

[17] Hiroshige Hirano, Toshiyuki Honda, Nobuyuki Moriwaki, Tetsuji Nakakuma, Atsuo Inoue, George Nakane, Shigeo Chaya, and Tatsumi Sumi, "2V/100-ns 1T/1C Nonvolatile Ferroelectric Memory Architecture with Bitline-Driven Read Scheme and Nonrelaxation Reference Cell", *IEEE Journal of Solid-State Circuits*, May 1997, Vol. 32, No. 5, pp.649-654.

[18] Betty Prince, "Emerging Memories: Technologies and Trends", Kluwer Academic Publishers © 2002.

[19] Linda Geppert, "The new indelible memories", *IEEE Spectrum*, March 2003, pp. 49-54.

[20] S.Kawashima *et al.*, "A Bit-Line GND sense technique for Low voltage operation FeRAM", *2001 IEEE Symposium On VLSI Circuits Digest of Technical Paper*, pp.127-128.

[21] Mun-Kyu Choi *et al.*, "A 0.25um 3.0-V 1T1C 32-Mb nonvolatile ferroelectric RAM with address transition detector and current forcing latch sense amplifier scheme", *IEEE Journal of Solid-State Circuits*, November 2002, Vol. 37, No. 11, pp. 1472 – 1478.

[22] Young Min Kang, Choong Heui Chung, Sang Hyun OH, Beelyong Yang, Seaung Suk LEE, Suk Kyoung Hong and Nam Soo Kang, "Characterization of Polarization Switching Behavior of Pt/SrBi2Ta2O9/Pt Ferroelectric Capacitors in Ferroelectric Random Access Memory", *Japanese Journal of Applied Physics*, 2002, Vol. 41, pp. 694–697.

[23] K.S.Tang, W.S.Lau, and G.S.Samudra, "Trends in DRAM dielectrics", *IEEE Circuits and Devices Magazine*, May 1997, Vol.13, No.3, pp. 27 – 34.

[24] "FRAM Guide Book", www.fujitsu.com, January 2005.

[25] Kiyoo Itoh, "Vlsi Memory Chip Design", Springer series publication 2001,Germany.

[26] Raffaele Zambrano, and Stradale Primosole, "Applications and issues for ferroelectric NVMs", *Materials Science in Semiconductor Processing 5* (2003) 305–310.

[27] M.Bauer *et al.*, "A multilevel-Cell 32Mb Flash Memory", *1995 IEEE International Solid-State Circuits Conference*, pp. 132-133.

[28] M.Aoki *et al.*, "A 16-Level/Cell Dynamic Memory," *IEEE Journal of Solid-State Circuits*, April 1987, Vol. SC-22, No.2, pp. 297-299.

[29] Bruno Ricco *et al.*, "Nonvolatile Multilevel Memories for digital applications", *Proceedings of the IEEE*, December 1998, Vol.86, No.12, pp.2399-2421.

[30] T. Furuyama, T. Ohsawa, Y. Nagahama, H. Tanaka, Y. Watanabe, T. Kimura, and K. Muraoka. "An experimental 2-bit/cell storage DRAM for macrocell or memory-on-logic application", *IEEE Journal of Solid-State Circuits*, April 1989, Vol.24, No.2, pp. 388–393.

[31] P. Gillingham, "A sense and restore technique for multilevel DRAM", *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, July 1996, Vol.43, No.7, pp.483–486.

[32] G.Birk, "Evaluation, design and implementation of multilevel DRAM", M.Sc. Thesis, University of Alberta, AB, Canada, 1999.

[33] Alexander Bugeja, "High-speed, high-precision digital-analog converters designed for spectral performance", PhD Thesis, University of Illinois at Urbana-Champaign, 2000.

[34] Katayoun Falakshahi, "High-speed high-resolution D/A conversion in CMOS", PhD thesis, Stanford University, 1999.

[35] Phillip E. Allen and Douglas R. Holberg, "CMOS Analog Circuit Design", 2nd Edition, Oxford University Press, 2002.

[36] K.Ola Andersson and J.Jacob Wikner, "Characterization of a CMOS Current-Steering DAC using State-Space Models", *Proceedings of 43rd IEEE Midwest Symposium on Circuits and Systems*, Lansing ML, Aug 8-11, 2000.

[37] Byung-Gil Jeon *et al.*, "A novel cell charge evaluation scheme and test method for 4Mb non-volatile ferroelectric RAM", *Proceedings of 6th International Conference on VLSI and CAD*, October 1999, pp. 281 – 284.

[38] S.Y.Lee and Kinam Kim, "Prospects of emerging new memory technologies" *Proceedings of IEEE International Conference on Integrated Circuit Technology, 2004*, pp. 45-51.

[39] Byung-Gil Jeon, Mun-Kyu Choi, Yoonjong Song, Seung-Kyu Oh, Yeonbae Chung, Kang-Deog Suh, and Kinam Kim, "A 0.4-m 3.3-V 1T1C 4-Mb Nonvolatile Ferroelectric RAM with Fixed Bitline Reference Voltage Scheme and Data Protection Circuit", *IEEE Journal Of Solid-state Circuits*, November 2000, Vol.35, No.11, pp. 1690-1694.

[40] Ali Sheikholeslami, "Circuit Design and Modeling of Ferroelectric Memories", PhD Thesis, University of Toronto, ON, Canada, 2000.

[41] FM25L256, "256Kb FRAM Serial 3V Memory", Ramtron product data sheets.

[42] Behzad Razavi, "Design of Analog CMOS Integrated Circuits" Boston, MA : McGraw-Hill, 2001.

[43] I.Olkin, L.Gleser, C.Derman, "Probability Models and Applications", Macmillan publishing, New York, 1980.

[44] Kevin Hawley, "Nonvolatile multilevel memories", *The IEEE Computer Society's Student Newsletter*, Summer 1999 Vol. 7, No. 2.

[45] J.F.Scott, "Ferroelectric Memories", Springer, New York, 2000.

[46] Gary F. Derbenwick and Stephen C. Philpy, "A talk on Ferroelectric Memory Reliability and Qualification", Celis Semiconductor Corporation, December, 11-12, 2001, Pasadena, CA.

[47] J.A.Rodriguez *et al.*, "Reliability Properties of Low-Voltage Ferroelectric Capacitors and Memory Arrays", *IEEE Transactions on Device and Materials Reliability*, September 2004, Vol.4, No.3, pp. 436-449.

[48] G.Torelli, C.Calligaro, A.Manstretta, A.Pierin, and P.Rolandi, "Charge-and-split sense amplifier for multilevel nonvolatile memories", *Electronics Letters*, 13 May 1999, Vol. 35, No.10, pp.796 – 798.

[49] D.Montanari, J.Van Houdt, G.Groeseneken, H.E.Maes, "Novel level-identifying circuit for flash multilevel memories", *IEEE Journal of Solid-State Circuits*, July 1998, Vol.33, No.7, pp.1090 – 1095.

[50] Yun Wu and Guozhong Cao, "Ferroelectric and dielectric properties of strontium bismuth niobate vanadates", *Journal of Materials Research*, July 2000, Vol.15, No.7, pp. 1583-1590.

[51] Brent Keeth, Jacob Baker, "DRAM Circuit Design", IEEE press New York 2000.

[52] F.Bedeschi *et al.*, "4-Mb MOSFET-Selected Phase-Change Memory Experimental Chip", *Proceeding of the 30th European Solid-State Circuits Conference*, 2004, pp. 207 – 210.

[53] J.DeBrosse et.al., "A 16Mb MRAM featuring bootstrapped write drivers", *2004 Symposium on VLSI Circuits Digest of Technical Papers*, pp. 454 – 457.

Appendix A

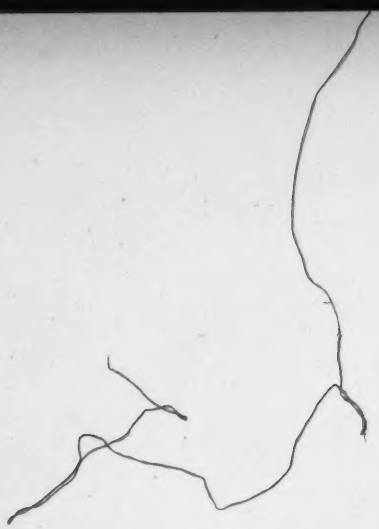
A.1

```
Verilog AMS model of Operational Amplifier
//
// -Operational amplifier
//
// vin_p,vin_n:          differential input voltage [V,A]
// vout:      output voltage [V,A]
// vref:      reference voltage [V,A]
// vspply_p:             positive supply voltage [V,A]
// vspply_n:             negative supply voltage [V,A]
//
// INSTANCE parameters
//   gain          = gain []
//   freq_unitygain = unity gain frequency [Hz]
//   rin           = input resistance [Ohms]
//   vin_offset    = input offset voltage referred to negative [V]
//   ibias         = input current [A]
//   iin_max       = maximum current [A]
//   slew_rate     = slew rate [A/F]
//   rout          = output resistance [Ohms]
//   vsoft         = soft output limiting value [V]
//
// MODEL parameters
//   {none}
//

module opamp(vout, vref, vin_p, vin_n, vspply_p, vspply_n);
input vref, vspply_p, vspply_n;
inout vout, vin_p, vin_n;
electrical vout, vref, vin_p, vin_n, vspply_p, vspply_n;
parameter real gain = 835e3;
parameter real freq_unitygain  = 1.0e6;
parameter real rin = 1e6;
parameter real vin_offset = 0.0;
parameter real ibias = 0.0;
parameter real iin_max = 100e-6;
parameter real slew_rate = 0.5e6;
parameter real rout = 80;
parameter real vsoft = 0.2;
  real c1;
  real gm_nom;
  real r1;
  real vmax_in;
  real vin_val;

  electrical cout;


  analog begin

    @( initial_step or initial_step("dc") ) begin
          c1 = iin_max/(slew_rate);
          gm_nom = 2 * `PI * freq_unitygain * c1;
          r1 = gain/gm_nom;
          vmax_in = iin_max/gm_nom;
    end

    vin_val = V(vin_p,vin_n) + vin_offset;

    //
    // Input stage.
```

```
    //
I(vin_p, vin_n) <+ (V(vin_p, vin_n) + vin_offset)/ rin;
    I(vref, vin_p) <+ ibias;
    I(vref, vin_n) <+ ibias;

    //
    // GM stage with slewing
    //
    I(vref, cout) <+ V(vref, cout)/100e6;

    if (vin_val > vmax_in)
      I(vref, cout) <+ iin_max;
    else if (vin_val < -vmax_in)
      I(vref, cout) <+ -iin_max;
    else
      I(vref, cout) <+ gm_nom*vin_val ;

    //
    // Dominant Pole
    //
    I(cout, vref) <+ ddt(c1*V(cout, vref));
    I(cout, vref) <+ V(cout, vref)/r1;

    //
    // Output Stage
    //
    I(vref, vout) <+ V(cout, vref)/rout;
    I(vout, vref) <+ V(vout, vref)/rout;

    //
    // Soft Output Limiting
    //
    if (V(vout) > (V(vspply_p) - vsoft))
      I(cout, vref) <+ gm_nom*(V(vout, vspply_p)+vsoft);
    else if (V(vout) < (V(vspply_n) + vsoft))
      I(cout, vref) <+ gm_nom*(V(vout, vspply_n)-vsoft);
  end
endmodule
```

Pages 123-126 contained the description of the Ferroelectric Capacitor model and its AHDL implementation based on the Preisach theory of the hysteresis loop. The model was obtained from [3]. These pages have been removed because of the Copyright issues.

Pages 123-126 contained the description of the Ferroelectric Capacitor model and its AHDL implementation based on the Preisach theory of the hysteresis loop. The model was obtained from [3]. These pages have been removed because of the Copyright issues.



Pages 123-126 contained the description of the Ferroelectric Capacitor model and its AHDL implementation based on the Preisach theory of the hysteresis loop. The model was obtained from [3]. These pages have been removed because of the Copyright issues.

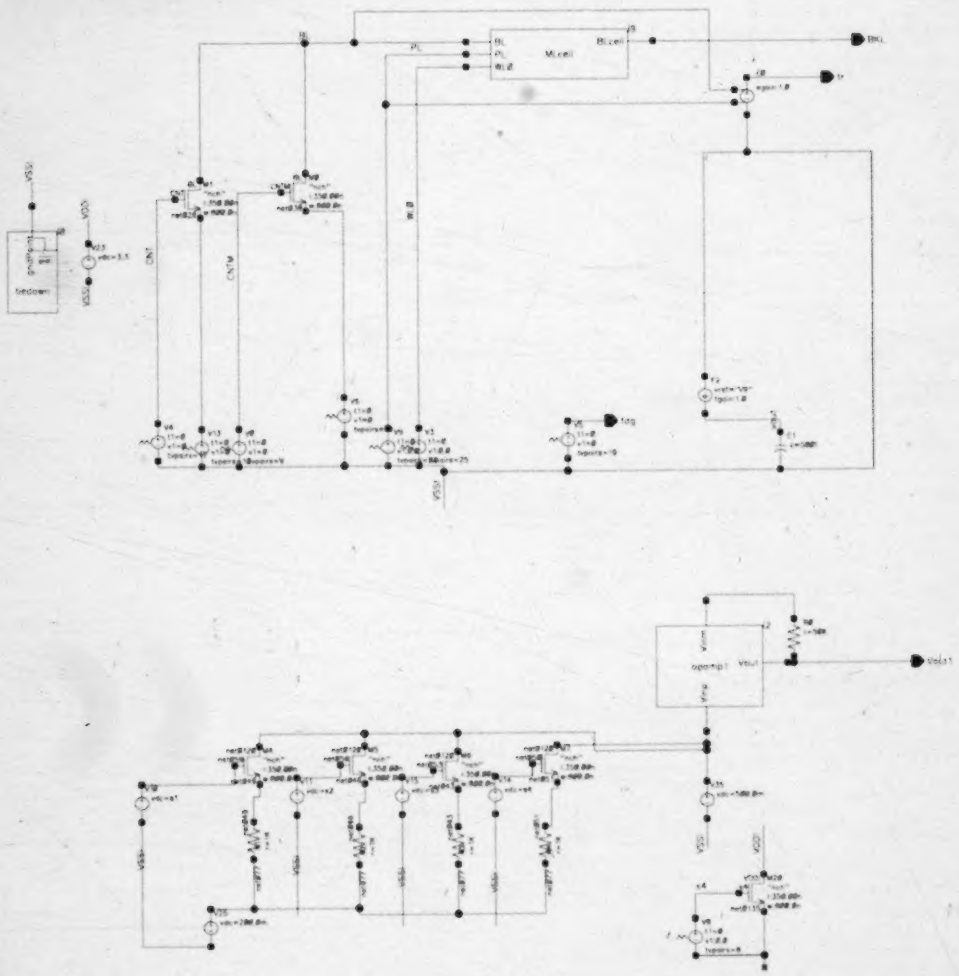
Pages 123-126 contained the description of the Ferroelectric Capacitor model and its AHDL implementation based on the Preisach theory of the hysteresis loop. The model was obtained from [3]. These pages have been removed because of the Copyright issues.

A.3

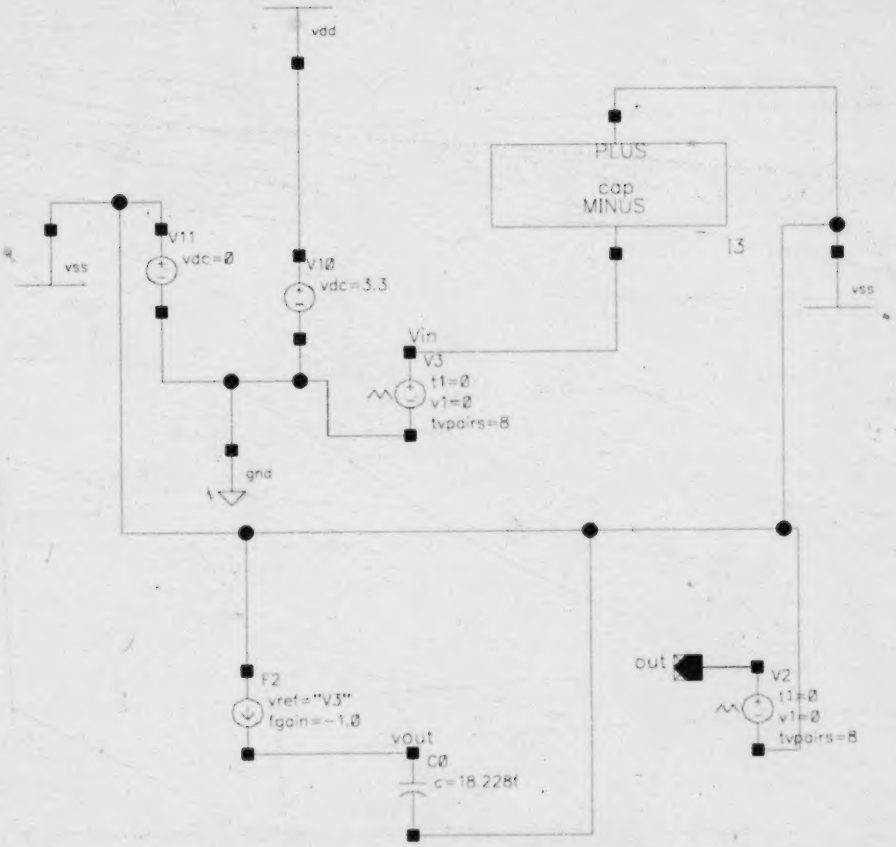
```
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
% Matlab Code to plot probability distribution curves
% Date 24 June 2004
%~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 % for +-20% variation in area .Total number of steps as 21
clear all;
Area=[]; % Vector for values of area
VL=[]; % Vector for values of readout voltages VL
VM=[]; % Vector for values of readout voltages VM
VH=[]; % Vector for values of readout voltages VH

VHVMVL=[VL,VM, VH];
N=60; % Number of points
y1=0;

  for i = 1:N,  % Since no zero indexing is allowed in a matrix in Matlab
    y1=y1 + Area(i), % Sum of all the values of Area
  end

u=y1/N;          %Find the Mean = Sum/No of values
v1=0;
for i = 1:N,
v1=v1 + (Area(i)-u)*(Area(i)-u); %Summation part for the calculation of sigma or say variance
end
d1=0.15*u % Standard deviation as percentage of mean.

%d1=sqrt(v1/N);  % Sigma is square root of variance and variance was V1/N

pi=3.14159;
e=2.718282;

  b1=(e.^-(((Area-u).^2)/(2*d1*d1)))/(sqrt(2*pi*d1*d1)); % calculation of pdf

%*****************************************************************
%Calculate Area of VL
A1=0;
for i = 1:N-1,
  A1=A1 + (0.5*(VL(i+1)-VL(i))*(b1(i)+b1(i+1))); % Area of Trapezoid 0.5*(base1+base2)*h
end

%Calculate Area of VM
A2=0;
for i = 1:N-1,
  A2=A2 + (0.5*(VM(i+1)-VM(i))*(b1(i)+b1(i+1)));
end
```

```
%Calculate Area of VH
A3=0;
for i = 1:N-1,
  A3=A3 + (0.5*(VH(i+1)-VH(i))*(b1(i)+b1(i+1)));
end

%*****************************************************************
%area(Area,b1);
b=[A1/A1*b1,A1/A2*b1, A1/A3*b1] % Multiply b1 with the ratio of areas b1*A2/A1(, b1*A2/A2,b2*A3/A1
%b=[(A1/(A1+A2+A3))*b1,(A2/(A1+A2+A3))*b1, (A3/(A1+A2+A3))*b1];
%plot(Area,b1);% This is the actual Normal Distribution of the statistical values of area
%hold;
%AXIS([400,2200,0.4,1.6]);
%plot(VHVMVL,b1); % This is direct plot mapping from area to VH VM VL with the same values of PDF
plot(VHVMVL,b);   %Mapping for normalized values of area to do comparison
%plot(Area,VH);
%area(VHVMVL,b);
% The conclusion is the value of b changes with the values of A, if A values are taken in 100s than it gives till 0.002 which
is what we want.
```

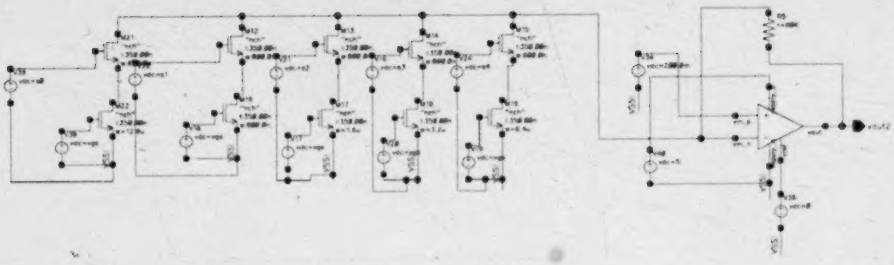
Appendix B



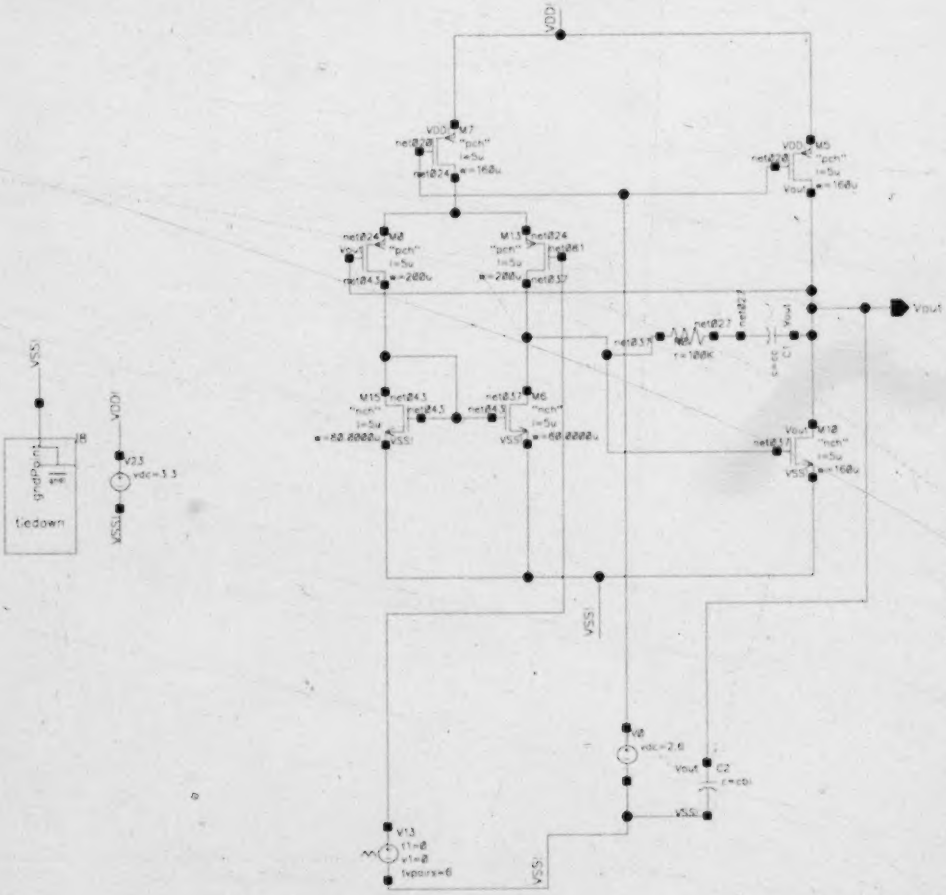
Appendix-B. 1 Schematic of MLsubloop



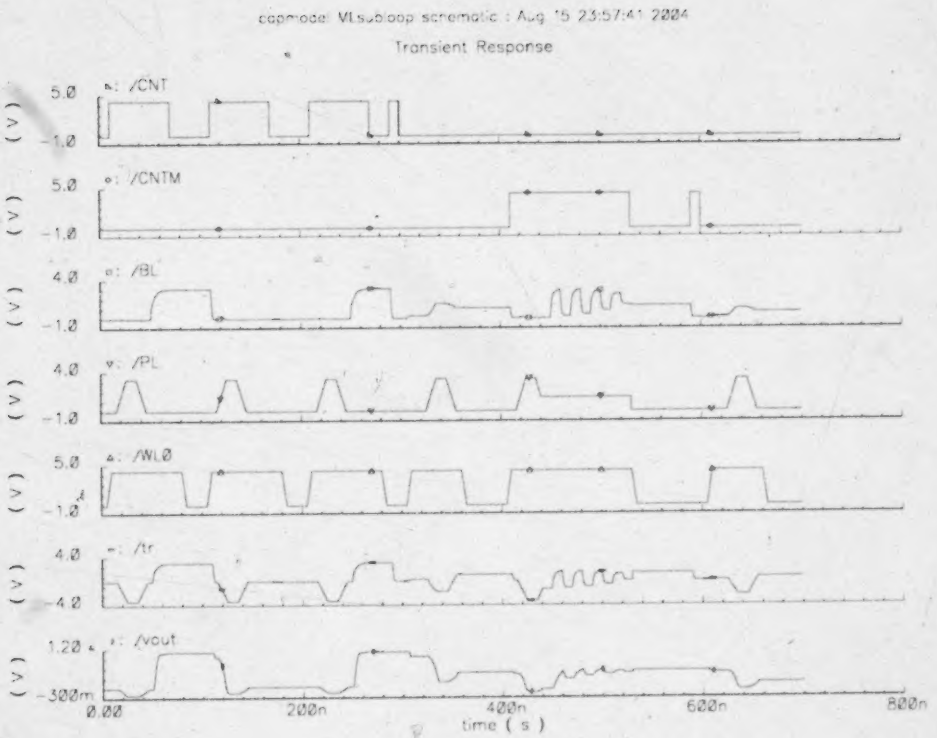
Appendix-B. 2 Schematic of hysteresis loop measurement circuit



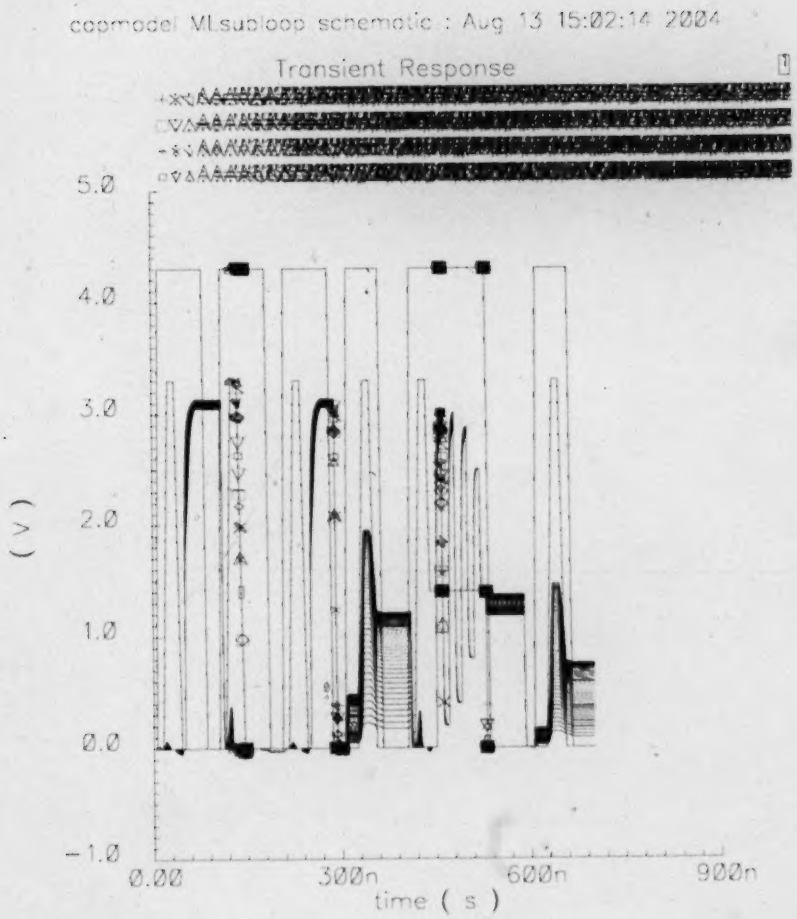
Appendix-B. 3 Schematic of DAC



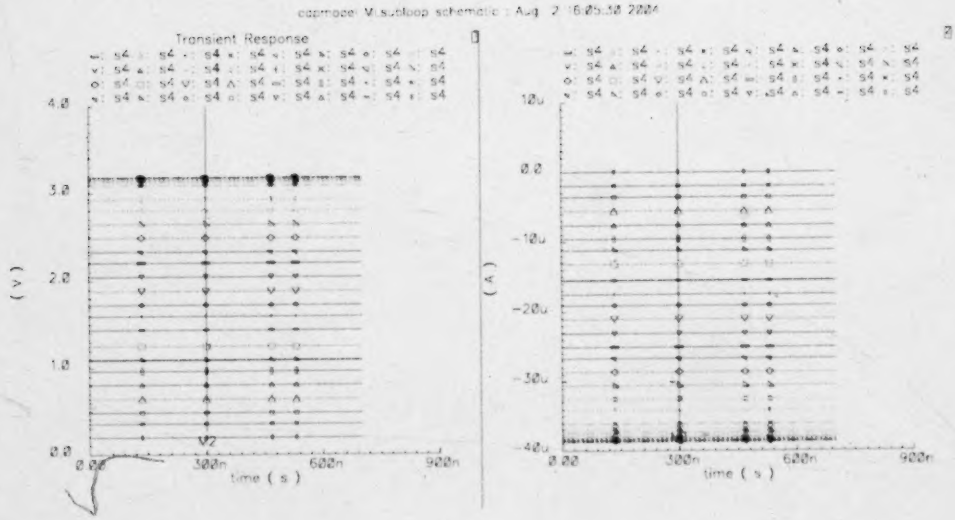
Appendix-B. 4 Schematic of Operational Amplifier



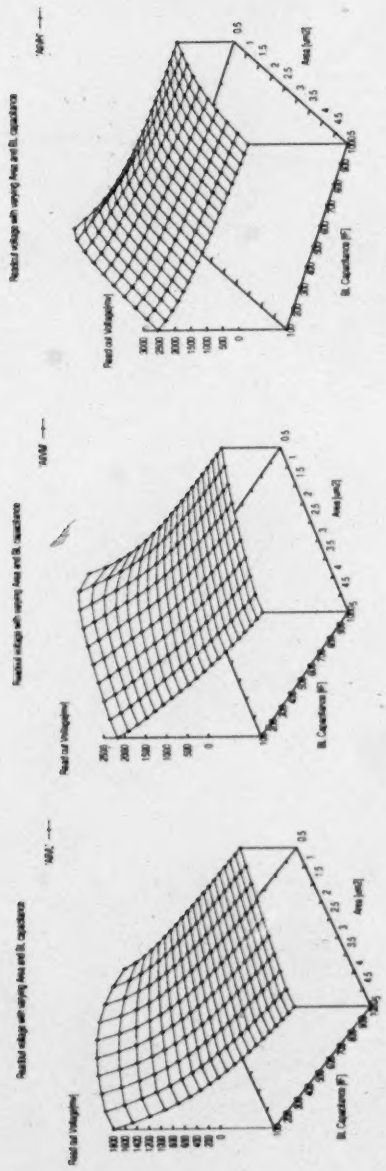
Appendix-B. 5 Read and Write waveforms



Appendix-B. 6 Parametric analysis for distribution plots



Appendix-B. 7 Appendix-B 1 Simulation Output Voltages and Current for 5-bit DAC



(c)

(b)

(a)

Appendix-B.8 Surface plots for (a) V_L (b) V_M (c) V_H with variation in capacitance and area

Appendix C

Area Estimation

The goal of the three-level FeRAM is to increase the density with the cell area remaining unaffected in the design process (means the FeCap area should be the same as that used for 1T1C). Sequential sensing can be used if read access time is not a major concern and area is a concern. For MLFeRAM to gain advantage over two-level the storage density for the chip should increase. However, due to the following extra circuitry there is an area overhead in MLFeRAM.

1.Double number of sense amplifiers on both sides for parallel sensing.

2.Precharge Block at the end of each column.

3.Autocalibration state machine

4.OPAMP for every 16 bitlines

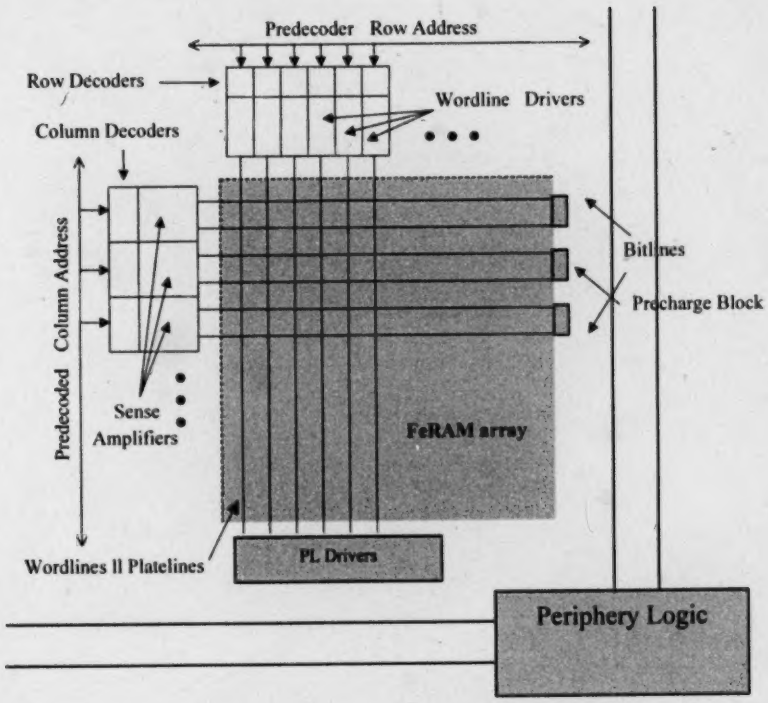
5.Thermometer decoding logic and A-to-D converter etc.

To have a rough estimate of the density we will do a comparison with a traditional two-level 8Mb FeRAM and 8Mb Chain FeRAM chip obtained from [5]. Area distribution for these two chips is shown below.

Table VIII FeRAM and Chain FeRAM Chip Size Comparisons

Two-Level 8Mbit FeRAM			**Chain FeRAM**		
SA and Column Decoders	(16.0 mm^2)	13.79%	SA and Column Decoders	(8.0mm^2)	8.79
Peri + Other	(14.6mm^2)	12.58%	Peri + Others	(15.3mm2)	16.81
XY Dec-Cross	(11.4mm^2)		--	--	--
Row Decoders	(8.0mm^2)	6.89%	Row Decoders	(10.0mm^2)	10.9
Plate Driver	(23.5mm^2)	20.2%	Plate Driver	(4.8mm^2	5.2
Memory Cell	(42.3mm^2)	36.4%	Memory Cell	(50.1mm^2)	54.94
Total	(116mm^2)	100%	Total	(91mm^2)	100%

Forming 8Mbit FeRAM as the basis of estimation we now evaluate the area for a three-level FeRAM. Figure below shows the major sections that a three-level FeRAM chip might have.



Block diagram of the Memory Array

Area due to SA

Assuming that we have a base array of 256Kbit arranged as 256 cells per bitline (column) and 1024 cells per wordline (row), we will have 32 such sections to make up 8Mb total array size. Also assuming that we are using parallel sensing the number of sense amplifiers will be doubled, as they will be needed at both the ends of the bitlines. However since the array organization in [5] is 128Kbit and we have 256Kbit the number of SAs will stay the same. Also assuming that the precharge block will be used at one of the ends of the columns and precharge block has roughly same area as sense amps, the new area accounted for SAs will be 1.5 times the previous one (conventional FeRAM); 16 x 1.5 = 24mm^2. This also includes area due to column decoders as it was in [5]. Since we are

accounting the area of column decoders as one and half times the previous value, we assume that the extra decoder logic required for thermometer coding is included in this, as well the area for PL decoders. The total number of sense amplifier will be 2048 (both sides) for each section hence 2048 x 32 = 65536 for all sections.

Area of Memory array cells

This area will remain the same as we are using the same 5.2μm^2 cell size, which is still quite big and our simulations support area sizes as small as 0.54μm^2 corresponding to capacitor sizes of 0.25μm^2. This closely matches with the cell sizes used in latest 0.13μm embedded FeRAM CMOS process.

Area due to plate line drivers.

This will remain the same as 23.5mm^2

Row decoders

This will remain the same as 8.0mm^2

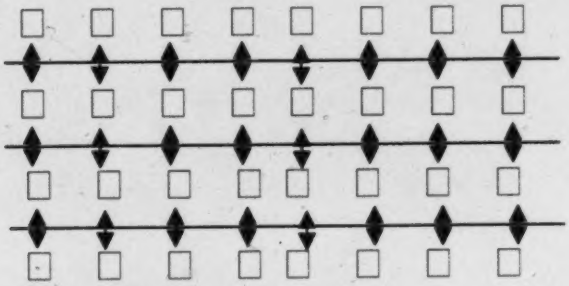
XY-decoder cross

This will remain the same as 11.4mm^2

Area due to middle level voltage driver OPAMP

Estimation: Lets assume that we are driving 16 bitlines with one driver. Which means that we will need 64 such drivers for each 256Kbit section and 64 x 8 such OPAMPs for the whole row of blocks arranged as 4 x 8. Now this OPAMP can be shared between two consecutive rows. Hence 3 such rows of OPAMPs. Which makes the total count of OPAMPs as 64 x 8 x 3 = 1536 for the whole 8Mb array see the fig below. The area of the OPAMP (Voltage Follower) used is about 5PMOS*160*2um^2 + 5NMOS*160*2 μm^2 + 500μm^2 (Cap CC) = 3700μm^2. So the total area is 3700*1536 = 5683200 μm^2. The calculations and the OPAMP designed was for 0.35 um however the area should remain the same even for 0.25

μm process (we are talking of this technology as the comparative example FeRAM [5] was implemented in this process) due to required sizes of the devices. A square indicates a 256Kbit section arranged in 256 rows and 1024 columns.



32 sections of 256 Kbit arranged as 4 x 8 to form 8192 Kbit

Area due Periphery and Others

This area will more or less remain the same. We assume that the area due to Finite state machine will be significantly less. Also in a conventional FeRAM since we would have needed extra area for generating V_{REF} voltage across every base array (of 256Kbit). We can counteractively neglect area overhead due to this finite state machine.

Hence the total area is:

Periphery + Others (14.6 mm^2)

XY Dec-Cross (11.4 mm^2)

SA + CD+ Precharge (24.0 mm^2)

RD (8.0 mm^2)

Plate Driver (23.5 mm^2)

V_M Driver (5.68mm^2)

Memory Cell (42.3 mm^2)

Total Area = 129.48 =~130 mm^2

However the total memory stored now is increased to12Mb instead of 8Mbit. Hence the memory density achieved now will be 12Mb/ 130 mm^2 = 9.23 cm^2. Which is 2.34 Mb/cm^2 improvement in the area density in Mb/cm^2. Which is 2.34/6.89= 33.96% equivalent decrease in area.

For 2-level FeRAM memory Density in Mb/ cm^2= 8Mb/116 mm^2= 6.8Mb/ cm^2.
For the chain FeRAM proposed in [5]=8Mb/76mm^2= 10.52 Mb/cm^2. Chain FeRAM achieves improvement of about 22-35% as shown in [5]

The improvement in area density for a three-level FeRAM chip will depend on several architectural features. For example a traditional reference generation scheme can be employed if area is a major concern. As well serial sensing (which uses a smaller number of sense amplifiers) could be used to decrease the sense amplifier area.